DOCUMENT RESUME

ED 104 930 TM 004 386

AUTHOR	Weiss, David J.
TITLE	Strategies of Adaptive Ability Measurement.
INSTITUTION	Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO	RR-74-5
PUB DATE	Dec 74
NOTE	91p.
EDRS PRICE	MF-$0.76 HC-$4.43 PLUS POSTAGE
DESCRIPTORS	*Academic Ability; Achievement Tests; Aptitude Tests; Electronic Data Processing; *Individual Differences; Low Ability Students; Measurement Techniques; Scoring Formulas; *Test Construction; *Testing; Testing Problems; Tests; Test Wiseness
IDENTIFIERS	*Adaptive Ability Testing

ABSTRACT

A number of strategies are described for adapting ability test items to individual differences in ability levels of testees. Each strategy consists of a different set of rules for selecting the sequence of test items to be administered to a given testee. Advantages and disadvantages of each strategy are discussed, and research issues unique to the strategy are described. Strategies reviewed are differentiated into two-stage approaches and multi-stage approaches. Several variations of the two-stage approach are described. Multi-stage strategies include fixed branching and variable branching strategies. Fixed branching strategies reviewed include a number of variations of the pyramidal approach (e.g., constant step size pyramids, decreasing step size pyramids, truncated pyramids, multiple-item pyramids), the flexilevel test, and the stradaptive test. Variable branching approaches include two Bayesian strategies and two maximum likelihood strategies. The various strategies are compared with each other on important characteristics and on practical considerations, and ranked on their apparent potential for providing measurement of equal precision at all levels of ability. (Author)

ED104930

TM

STRATEGIES OF ADAPTIVE ABILITY MEASUREMENT

David J. Weiss

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Research Report 74-5

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

December 1974

Prepared under contract No. N00014-67-A-0113-0029
NR No. 150-343, with the Personnel and
Training Research Programs, Psychological Sciences Division
Office of Naval Research

Approved for public release; distribution unlimited.
Reproduction in whole or in part is permitted for
any purpose of the United States Government.

TM 004 386

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 74-5	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Strategies of Adaptive Ability Measurement	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): David J. Weiss	8. CONTRACT OR GRANT NUMBER(s): N00014-67-0113-0029
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology University of Minnesota Minneapolis, Minnesota 55455	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: PE.:61153N PROJ.:RR042-04 T.A.:RR042-04-01 W.U.:NR150-343
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217	12. REPORT DATE: December 1974 13. NUMBER OF PAGES: 78
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

testing	sequential testing	programmed testing
ability testing	branched testing	response-contingent testing
computerized testing	individualized testing	automated testing
adaptive testing	tailored testing	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A number of strategies are described for adapting ability test items to individual differences in ability levels of testees. Each strategy consists of a different set of rules for selecting the sequence of test items to be administered to a given testee. Advantages and disadvantages of each strategy are discussed, and research issues unique to the strategy are described. Strategies reviewed are differentiated into two-stage

approaches and multi-stage approaches. Several variations of the two-stage approach are described. Multi-stage strategies include fixed branching and variable branching strategies. Fixed branching strategies reviewed include a number of variations of the pyramidal approach (e.g., constant step size pyramids, decreasing step size pyramids, truncated pyramids, multiple-item pyramids), the flexilevel test, and the stradaptive test. Variable branching approaches include two Bayesian strategies and two maximum likelihood strategies. The various strategies are compared with each other on important characteristics and on practical considerations, and ranked on their apparent potential for providing measurement of equal precision at all levels of ability.

Contents

Introduction..... 1

Two-stage Strategies..... 3
Scoring..... 7
Advantages and Limitations..... 9
Research Issues..... 10

Multi-stage Strategies..... 12
Fixed Branching Models..... 12
Pyramidal models..... 12
Constant step size pyramids..... 12
Decreasing step size pyramids..... 18
Truncated pyramids..... 22
Multiple-item models..... 25
Differential response option branching..... 26
Scoring..... 30
Advantages and limitations..... 34
Research issues..... 36
The flexilevel test..... 36
Scoring..... 41
Advantages and limitations..... 41
Research issues..... 42
The stradaptive test..... 44
Scoring..... 50
Advantages and limitations..... 53
Research issues..... 54
Variable Branching Models..... 54
Bayesian strategies..... 55
Novick's strategy..... 56
Owen's strategy..... 58
Advantages and limitations..... 59
Research issues..... 61
Maximum likelihood strategies..... 62
Advantages and limitations..... 66
Research Issues..... 67

Evaluation..... 67
Bayesian and Maximum Likelihood Strategies..... 68
The Stradaptive Test..... 70
Truncated Pyramids..... 70
Two-stage Tests..... 71
Multiple-item Pyramids..... 72
Other Pyramidal Models..... 72
The Flexilevel Test..... 73
Summary..... 73

References..... 75

List of Figures

Figure		Page
1.	A two-stage strategy with peaked routing test	4
2.	A two-stage strategy with rectangular routing test	6
3.	A double-routing two-stage strategy	8
4.	A pyramidal item structure with constant step size	13
5.	Illustrative paths through a constant step size pyramid	15
6.	Paths through a constant step size pyramid with up-one/down-two branching rule	17
7.	Item structure for a decreasing step size pyramidal test	19
8.	Pyramidal item structure for a six-stage Robbins-Monro shrinking step size pyramidal test	21
9.	Truncated pyramidal item structures with reflecting and retaining barriers	24
10.	A three-items-per-stage pyramidal test structure	27
11.	A pyramidal item structure using differential response option branching	29
12.	Pyramidal scoring based on two methods of using final response data	32
13.	Item structure for a ten-stage flexilevel test	38
14.	Sample paths through a ten-stage flexilevel test	39
15.	Distribution of items, by difficulty level, in a stradaptive test	45
16.	Report on a stradaptive test for a consistent testee	48
17.	Report on a stradaptive test for an inconsistent testee	49
18.	A diagrammatic representation of Novick's Bayesian testing strategy	57
19.	Report on a Bayesian test	60
20.	Hypothetical response records from Urry's maximum likelihood adaptive testing strategy	63

STRATEGIES OF ADAPTIVE ABILITY MEASUREMENT

For almost sixty years, the predominant mode of administration of ability tests has been the paper and pencil multiple-choice test. These tests are usually administered to testees in groups and are designed to require all individuals in the group to answer all test items, whether or not they are appropriate for any given individual. If parts of a test are too difficult for an individual testee, he/she may experience frustration and thus react negatively to the test. When a test is too difficult, some people tend to guess, although there appear to be wide individual differences in the tendency to guess. Alternatively, parts of a test may be too easy for other individuals; in this case the testee may not be sufficiently challenged to put forth maximum effort or he may become bored and, therefore, also not respond in an optimal way. In both of these cases--a test which is too difficult for some people and too easy for others--the extraneous factors introduced in test responses may lower the accuracy of test scores, leading to erroneous decisions about the individual being tested.

In addition to these potential psychological effects of a fixed item pool on testees of different ability levels, psychometric theory indicates adverse effects on the reliability and validity of test scores. In a series of theoretical studies comparing conventional and tailored (adaptive) tests, Lord's (1970, 1971a,c,d,e) results indicate that a test score will most accurately reflect an individual's ability when the probability of an individual answering each item correctly is .50. Hick (1951) independently reached the same conclusion from developments in the field of information theory. Thus, these findings indicate that a test will have lower precision of measurement (i.e., higher standard error of measurement for a given test score; see Green, 1970, p. 186) when the probability of a correct response to an item by a given testee is greater or less than .50. Very difficult items are those with probability of a correct response (proportion correct) between .50 and zero, and very easy items are those with probabilities greater than .50, and possibly approaching 1.00. It is obvious, therefore, that tests composed of items too easy or too difficult for a given testee will not measure that testee's ability as accurately as will a test with items of median difficulty for that person. Lord's (1970, 1971a,c,d,e) results show that conventional tests yield scores which are considerably less precise at ability levels on either side of the group mean ability than they are when the test is matched with the testee's ability level. The result of this lower precision of measurement is lower overall reliability and, therefore, probably lower validity.

The use of time limits in group testing, which are imposed on the testee primarily for the convenience of test administrators, may also introduce error in test scores. Some people respond appropriately to time limit pressures, pacing themselves in order

to maximize their test scores; others (e.g. those from minority cultures) are likely not to respond to time limit pressures in the same way. As a result, time limits may differentially affect certain testees, introducing an unknown amount of error in their test scores.

Although group tests were partially designed to eliminate administrator effects, certain administrator-testee interactions have been shown to exist even under group conditions of test administration (Weiss & Betz, 1973). Thus, administrator differences (e.g., administrator's attitudes, race) can introduce error into the examinee's test scores even on group paper and pencil tests.

For these and other reasons (Weiss & Betz, 1973), it is appropriate to investigate whether methods of test administration other than the conventional paper-and-pencil test can improve the reliability and validity characteristics of ability test scores. Adaptive testing can provide the vehicle for these improvements in psychometric characteristics of ability test scores. The basic idea of adaptive testing is that the test items administered to a given individual are selected to be appropriate to his ability level rather than to the average ability level of a group of individuals. The process of adapting item difficulties to the ability level of each individual is based on information obtained from each testee's responses to previous items on the test. The result is the selection of items with difficulties in the range of p=.50 for each individual, with consequent ability estimates which should be of relatively equal precision throughout the ability range.

Under certain adaptive testing strategies, the number of test items above a person's ability level can be minimized, thus possibly reducing frustration and guessing effects. Simultaneously, the number of items below each testee's ability level can be minimized, thus reducing the potential effects of boredom and conserving valuable testing time. Adaptive tests are generally untimed, permitting each testee to proceed at his own rate of speed. At the same time, the testee's response times can be measured providing data of potential utility for the psychologist. The number of test items administered to each person can also be individualized, thus drastically reducing testing time in some cases. Research to date on adaptive testing (Weiss & Betz, 1973) shows that it has considerable promise for greatly reducing testing time without reducing reliability and validity. Some studies have shown increases in reliability and validity under adaptive testing strategies even with sharply reduced numbers of items in the adaptive test, as compared to conventional paper and pencil tests.

In previous research on adaptive testing, tests have been administered by paper and pencil, by testing machines, and by computer. In computerized adaptive testing, ability tests may be re-designed for administration to each testee on cathode-ray-

typewriter (CRT) terminals or slide projector screens connected to an on-line computer system (e.g., Cory, 1974; DeWitt & Weiss, 1974). After a test item is presented on the CRT or projector screen, the testee may respond to the test question on a typewriter keyboard, or by a light-pen on the CRT screen. The response to each test question is immediately scored by the computer, the next test question is chosen by the computer program according to a specified adaptive testing strategy, and that item is presented for the testee's response.

All strategies of adaptive testing operate from a set of pre-normed test items, or an item "pool" (e.g., McBride & Weiss, 1974). The strategies differ in the way in which the items in the pool are chosen for adaptation to individual differences in ability. Thus, different strategies represent different ways of moving a testee through an item pool by some sequential procedure.

The objective of the present paper is to describe the various strategies of adaptive testing which have been proposed and to evaluate and compare their characteristics on logical grounds.[1] The two general classes of procedures proposed to date include two-stage strategies and multi-stage strategies; the latter can be further divided into fixed-branching models and variable branching models.

TWO-STAGE STRATEGIES

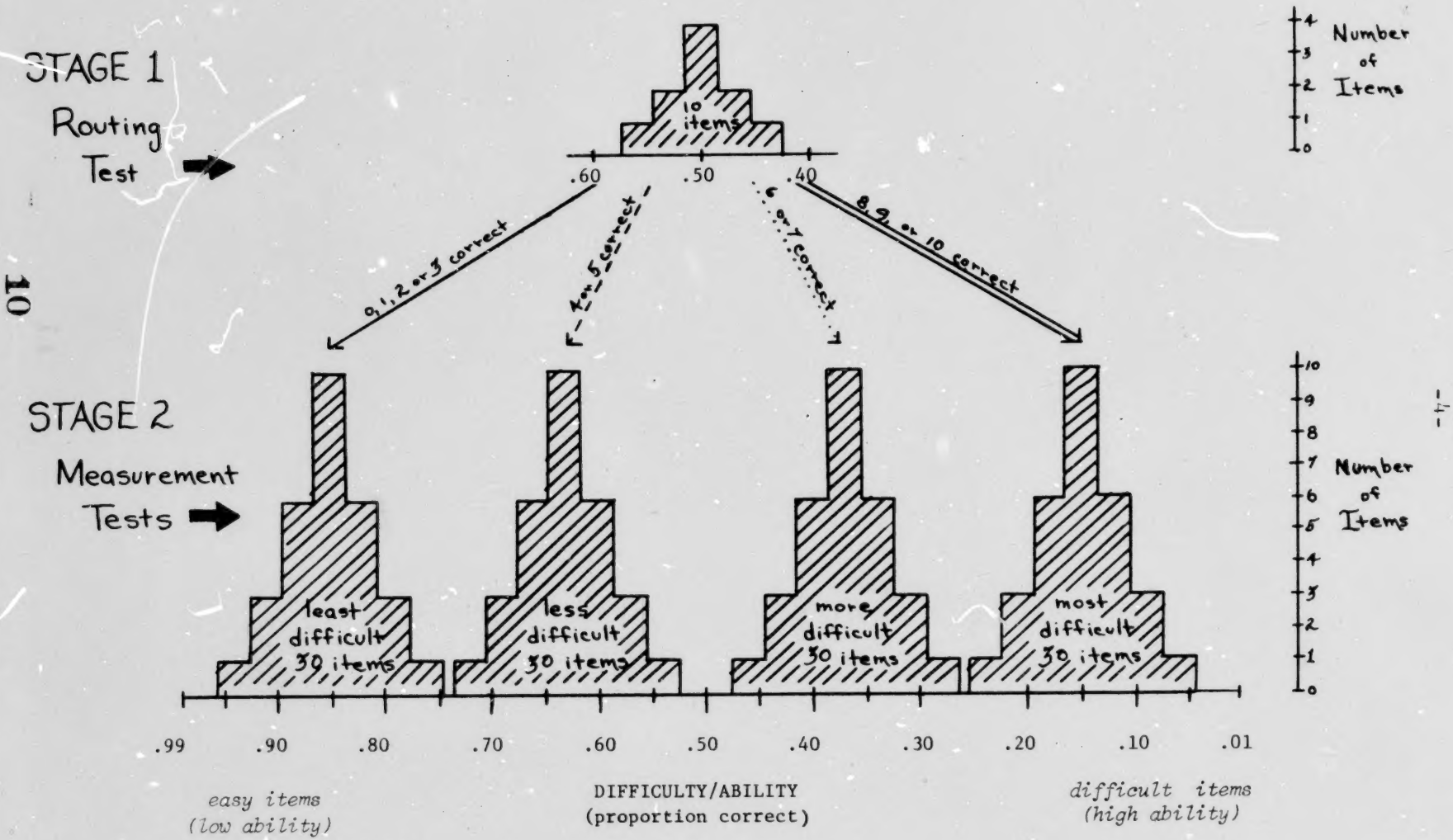
The two-stage test is the simplest of the adaptive testing strategies. Two-stage strategies have been studied by Angoff and Huddleston (1958), Betz and Weiss (1973, 1974), Cleary, Linn, and Rock (1968; Linn, Rock & Cleary, 1969), Lord (1971e), and Wood (1971). This strategy usually consists of a routing test and a measurement test. The routing test may be a broad-range ability test composed of items of differing difficulties, varying from very easy items to very difficult items; it can also be a "peaked" ability test in which all test items are at the average difficulty level for the group to be tested. The routing test is typically a short test designed to provide an initial estimate of an individual's ability level.

Based on his score on the routing test, each testee is branched to one of a number of measurement tests. Each measurement test is peaked at a different level of difficulty and is designed to differentiate among the abilities of individuals within a narrower range of ability than the routing test.

Figure 1 is a diagram of a hypothetical two-stage strategy

[1]Technical data derived from applications of these adaptive testing strategies are reviewed in detail by Weiss and Betz (1973).

Figure 1
A Two-Stage Strategy with Peaked Routing Test



using a peaked routing test. Tests of this general type were studied by Angoff and Huddleston (1958), Betz and Weiss (1973, 1974), Lord (1971e), and Wood (1971). In Figure 1 the routing test consists of 10 items, all answered correctly by from 45% to 55% of the group on which the items were normed. If each test item is scored correct or incorrect, scores on the routing test can vary from 0 to 10. Branching, or routing to the second stage test, occurs on the basis of scores on the stage 1 routing test. Testees who score 0, 1, 2 or 3 on the routing test are assumed to be of lowest ability and are branched, following the solid line, to the least difficult measurement test. That test is composed of 30 items clustered around a difficulty level of .85 (i.e., 85% of the norming group answered those items correctly). In a similar fashion, testees who score 4 or 5 on the routing test are branched to a measurement test of greater difficulty, with items clustered around a difficulty of .63. Branching for scores of 6 or 7 is to a more difficult measurement test (average p=.37), while testees with scores of 8,9 or 10 on the routing test are branched to the most difficult measurement test, with items answered correctly on the average by only 15% of the norming group.

The time savings involved in two-stage adaptive testing can be seen from an examination of Figure 1. If all the test items shown in Figure 1 were administered to each testee, each person would complete 130 items (i.e., 10 items from the routing test and 30 items from each of the four measurement tests). For each person, however, at least 90 of the items would be either too easy or too difficult. Using the two-stage strategy each testee would complete only 40 items. If the routing test branches a person appropriately, of the 40 items administered to each testee, at least 30 (those in the measurement test) will be of approximately appropriate difficulty for that testee. The result is a test which is at least partially adapted to each testee's level of ability. Such adaptation to individual differences in ability should increase test-taking motivation and should have positive effects on reliability and validity of measurement (Weiss & Betz, 1973). Betz and Weiss (1974) have shown such a two-stage test to have higher reliability (test-retest) than a conventional test of the same length, when the potential for memory effects was equal.

A number of variations have been proposed for two-stage testing strategies. Figure 2 shows a two-stage testing strategy with a rectangular distribution of items in the routing test. Thus, rather than having all items in the routing test clustered around a mean value for a specified group, there is only one item at each of a number of levels of difficulty, spanning the full range of difficulties. Routing to measurement tests still occurs on the basis of score on the routing test. The two-stage strategy in Figure 2 has five measurement tests placed at constant intervals on the difficulty continuum. One measurement test is at the median of the difficulty/ability scale, at the same average difficulty level as the peaked routing test used in the

Figure 2
A Two-Stage Strategy with Rectangular Routing Test

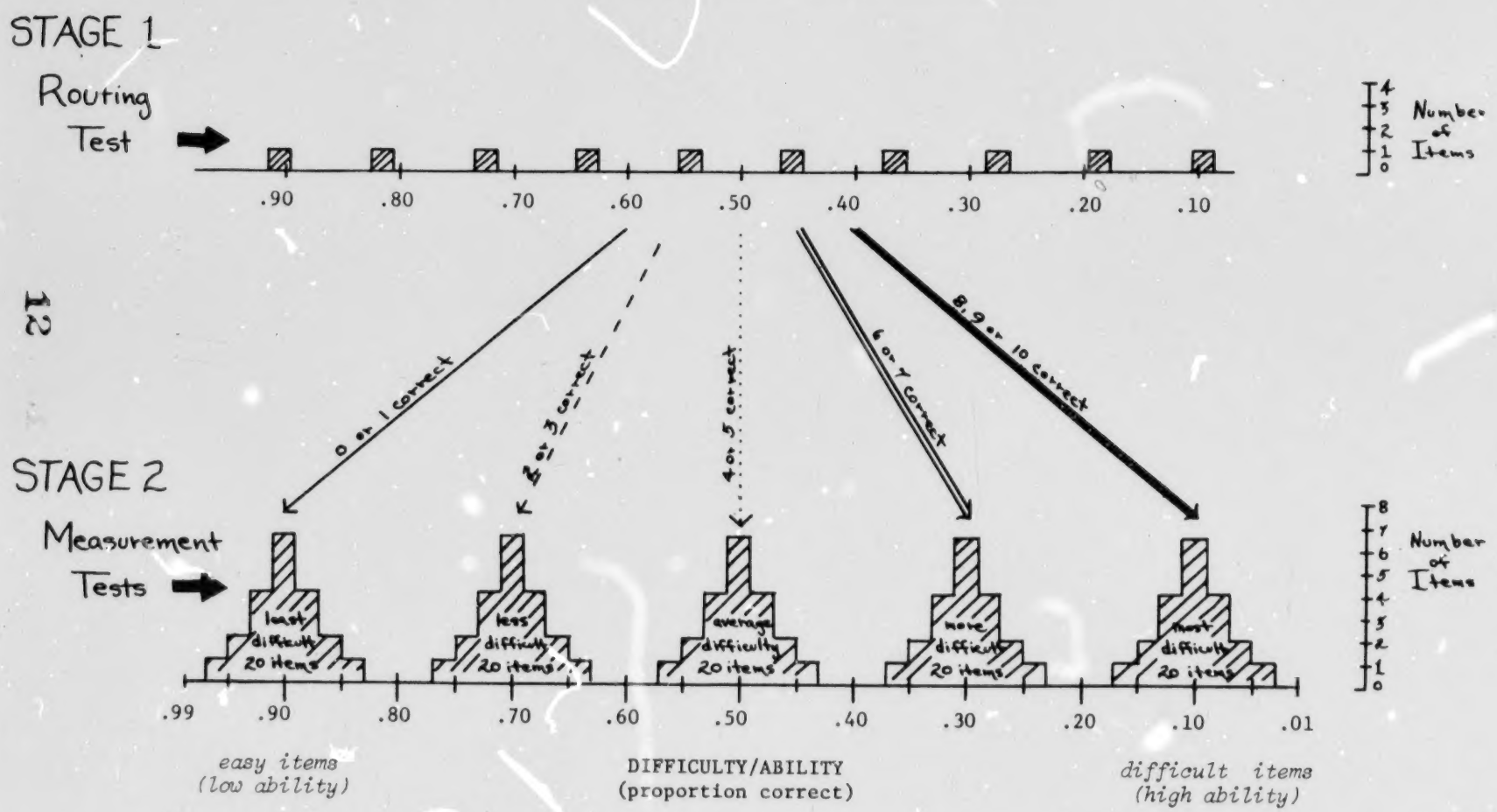

STAGE 1
Routing Test
.90 .80 .70 .60 .50 .40 .30 .20 .10
Number of Items
0 or 1 correct
2 or 3 correct
4 or 5 correct
6 or 7 correct
8, 9 or 10 correct
STAGE 2
Measurement Tests
least difficult 20 items
less difficult 20 items
average difficulty 20 items
more difficult 20 items
most difficult 20 items
.99 .90 .80 .70 .60 .50 .40 .30 .20 .10 .01
Number of Items
easy items (low ability)
DIFFICULTY/ABILITY (proportion correct)
difficult items (high ability)

Figure 1 strategy. A two-stage test using a rectangularly distributed routing test was studied by Cleary, Linn and Rock (1968) and Linn, Rock and Cleary (1969).

Cleary, et al. (1968) and Linn, et al. (1969) report on several other variations of the basic two-stage strategy. In each case, the variation involves changes in the routing test. One of their variations was a double routing test (see Figure 3). Their first routing test consisted of 10 items; scores on that test were dichotomized and used to route testees to two different second routing tests. The second routing tests were both peaked tests, scores on each of which were dichotomized to yield a final four-category classification as the result of routing. As Figure 3 shows, these authors used four 20-item peaked measurement tests. Thus, each testee answered 40 items, only 10 of which were answered by all testees.

The same authors also studied two other variations of a two-stage routing test. In their "group discrimination" method the routing test was constructed using items whose difficulties were found to be significantly different for groups classified by total scores on a parent conventional test. Their second method used a sequential routing test. In that test, items were presented one at a time and, after each item, the probability that the testee was a member of each of four criterion groups was determined. Criterion groups were based on total score intervals on a conventional test. At some point in the sequential procedure, the testee was classified into one of the four criterion groups, and his classification determined the level of difficulty of the measurement test administered. These latter two methods of routing--group discrimination and sequential routing--appear not to be of general interest for the future development of two-stage models, however, because of their heavy dependence on total scores from a parent conventional test for their development.

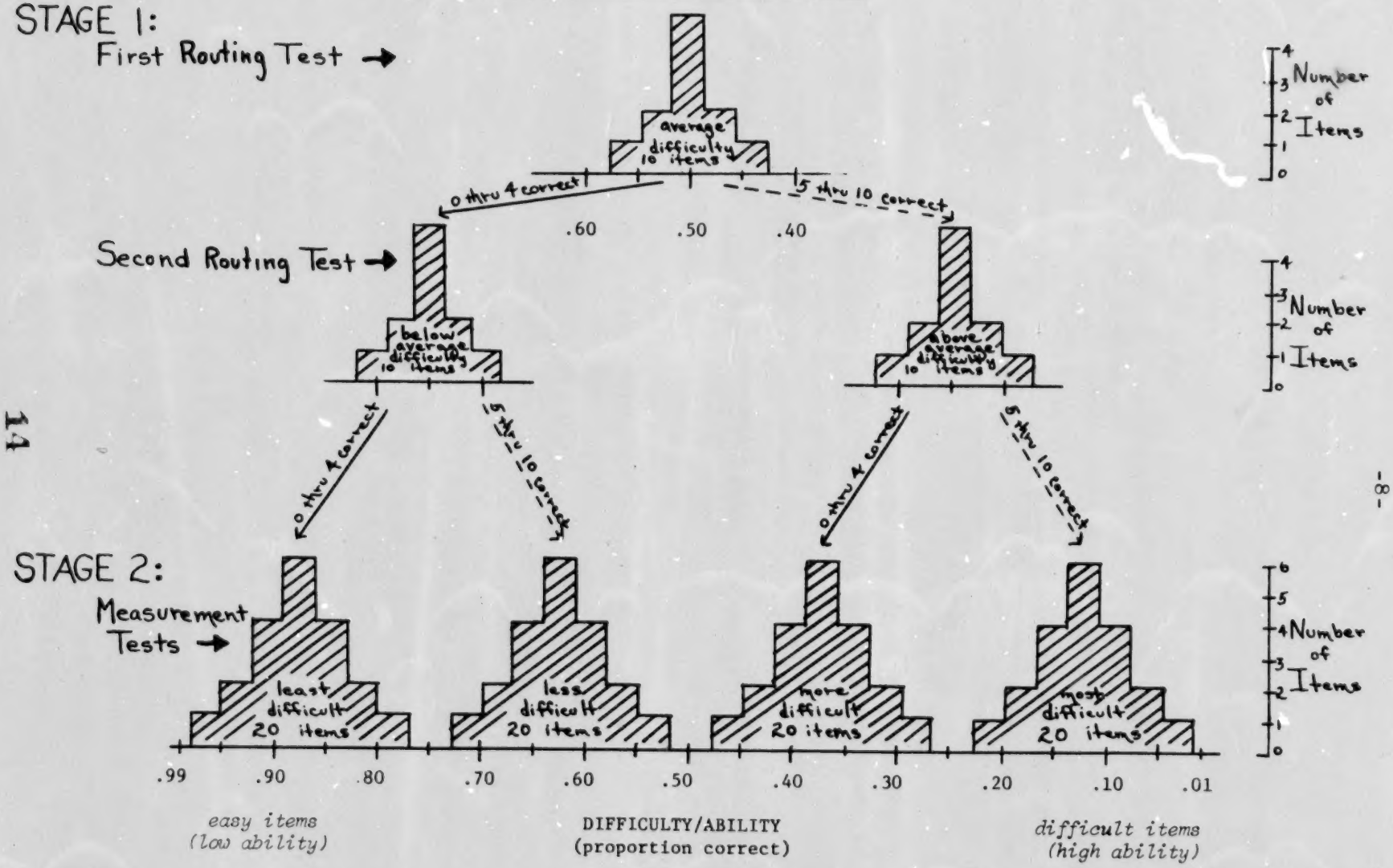
Scoring

Since different individuals take measurement tests composed of items of different difficulty levels, the number correct score commonly used in conventional tests is inappropriate for two-stage tests. Thus, new methods of scoring have been developed.

The average difficulty score is one method of scoring two-stage tests. This method consists of computing the average of the difficulties of all items answered correctly by the testees. Average difficulty scores can be based only on items in the measurement tests taken by a testee, or they can also include the items in the routing test.

Lord (1971e) developed a maximum likelihood procedure for estimating ability from responses to a two-stage test. His

Figure 3
A Double-Routing Two-Stage Strategy



formulas provide ability estimates in normal deviate form (mean=0, standard deviation=1) and the sampling variance of the ability estimates. Lord's formulas are based on the assumption that both the routing tests and measurement tests are peaked. The scoring formula uses the number correct on the routing and measurement tests for each testee relative to the number of items in the test, the chance score value, the difficulty of the peaked test, and the discriminations of the items (which are assumed to be equal for all items). These values are used to determine separate ability estimates for the routing and measurement tests. The final ability estimate is determined by combining the two separate ability estimates weighted inversely by their estimated variances.

Although Lord proposed weighting the ability estimates in this manner, he admits it is not optimal. An application of this weighting method (Betz & Weiss, 1973) indicated that the ability estimates derived from this method did not follow certain logical expectations when real test items were used. Betz and Weiss studied a computer-administered two-stage test with 10 items in the routing test and four 30-item measurement tests. Use of Lord's weighting procedure led to some illogical results; for example, a testee who answered four items correctly on the routing test and six items correctly on the relevant measurement test received a higher ability estimate than testees who obtained the same score on the routing test, but answered more items correctly (7 through 12 items) on the same measurement test. This difficulty was corrected by weighting the routing and measurement test ability estimates by the number of items in the respective subtests, rather than their maximum likelihood variance estimates, as Lord had suggested.

Cleary, et al. (1968, and Linn, et al., 1969) developed other methods for scoring their two-stage tests. Their methods are not of general use, however, since they are based on regression procedures designed to estimate scores on a "parent test" from those on a shorter, two-stage test.

Advantages and Limitations

The obvious advantage of the two-stage test in comparison to the conventional paper-and-pencil test lies in its adaptiveness. Although the two-stage strategy can be conceptualized as two conventional tests, in two-stage testing the routing test is scored before the measurement test is given. Thus, there is information on a testee's ability level which is used to adapt the remainder of the testing process to his ability level. Since the routing test is usually relatively short in comparison to the measurement test, the measurement test will provide more information per item over more items, and thereby may serve to reduce the negative psychological and psychometric effects of a conventional routing test which is either too difficult or too easy for a given testee.

Two-stage models have one major advantage over most other adaptive strategies. Because they are generally based on the use of two conventional tests, they are amenable to paper and pencil testing. All that is required is an answer sheet for the routing test that is easily scored, either by self-scoring procedures or use of a simple hand-scoring stencil. Given a routing test score determined by either of these two methods the administrator could then consult a table which converts raw scores on the routing test to measurement test assignment, and give the testee the appropriate measurement test booklet. Alternatively, but probably resulting in more routing errors, the testee could score his own routing test and follow instructions to the appropriate measurement test within a larger test booklet. This latter procedure, however, would require a highly motivated testee. The use of paper and pencil administration as an advantage of two-stage tests disappears, however, if more complex routing tests are used, such as the sequential routing test, or the double-routing procedure.

The logic of the two-stage testing strategy has inherent in it two primary disadvantages. Its primary limitation is routing errors. Routing errors are errors in the assignment of measurement tests due to errors of measurement in the routing test. These errors occur primarily for individuals whose scores fall near the cutting scores established for assignment to different measurement tests. That they can be fairly substantial is shown by routing errors for as many as 20% of the testees in both Linn, et al.'s (1969), and Angoff and Huddleston's (1958) studies of two-stage testing; Betz and Weiss (1973), however, showed 4% to 5% routing errors in their computer-administered two-stage test. One method of eliminating routing errors is to administer the two-stage test by computer, identify probable mis-routings early in the measurement test response record, and re-route the testee to another, more appropriate, measurement test.

A second limitation of two-stage models concerns the number of items administered to the testees. Research with the variable branching methods of adaptive testing, e.g., Bayesian strategies (see below, and Weiss & Betz, 1973, pp. 36-38) and the stradaptive test (see below, and Weiss, 1973) seems to indicate that individuals differ in the number of items they require to achieve a desired degree of accuracy of measurement. Two-stage models, as proposed to date, require that all individuals answer all items on the two conventional tests that comprise the routing and measurement tests. Thus, the procedure does not adapt the number of items presented to individual differences in consistency of response.

Research Issues

In order to fully explore the potential of two-stage testing models, a number of research questions need to be answered. Among these are the following:

1. What the the optimal characteristics of the routing test? Should routing tests be peaked, rectuangular, or

polymodally distributed? Could sequential decision classification procedures (e.g., Cowden, 1946; Cronbach & Gleser, 1965; Moonan, 1950; Wald, 1947) be used to develop tailored item sequences in the routing test that would be more efficient or effective than a conventional fixed-item routing test?

2. If a fixed-item conventional routing test is to be used, how can the score distribution on the routing test be best used to eliminate errors in routing? Should errors of measurement on the routing test be taken into account in the routing decision? How can an optimal set of cutting scores on a routing test be best developed to minimize errors in classification (i.e., assignment to measurement tests) due to the routing procedure?

3. Is there any generally optimal number of items in the routing test and the measurement tests? What is the optimal ratio of number of routing test items to number of measurement test items? Lord's (1971e) analyses provide some answers to this question. However, his findings are quite tentative since they are derived under very restrictive assumptions and on the basis of theoretical analyses only. Lord's results need to be confirmed and extended by empirical studies.

4. Some errors in routing are likely to occur even under optimal classification rules. Thus, it is appropriate to study how errors of routing might be detected for a given testee early in the mis-assigned measurement test. Then, when a two-stage test is computer-administered it can be programmed to recognize such errors and correct them by re-routing the testee to a different measurement test.

5. How many measurement tests should there be, and how should they be distributed across the ability continuum? Again, Lord (1971e) has provided some very tentative answers to this question, but further research is needed.

6. Does double routing, such as studied by Linn, et al. (1969), have any value? Would more than two routing tests be even more valuable? If so, what is the optimal number?

7. What method of scoring two-stage tests is best for what purposes? Do some scoring methods optimize reliability while others primarily increase validity or utility in specific situations?

Answers to these questions, and others that these answers will raise, will eventually result in the development of optimal designs for two-stage testing strategies.

MULTI-STAGE STRATEGIES

Fixed Branching Models

The majority of research in adaptive testing has used fixed-branching multi-stage testing strategies (Weiss & Betz, 1973). These strategies differ from two-stage strategies in that the two-stage strategy generally requires only one branching decision (i.e., from routing to measurement test) while the multi-stage strategies require a series of branching decisions. In the typical multi-stage test, a branching decision occurs after the testee responds to each test item. The fixed-branching multi-stage models all operate from an item pool which is both calibrated and structured. The item pool is usually calibrated in terms of the difficulties and discriminations of the test items. Item pools can be structured in a variety of ways. Each different way of structuring the item pool defines a different strategy of adaptive testing. The fixed-branching multi-stage strategies use the same item pool structure for all individuals, but individuals move through the structure in different ways. A "branching rule" is specified prior to testing and this rule determines how an individual moves from an item at one stage of testing to an item at the next stage. The branching rule, in conjunction with information on whether the testee answered a given item correctly or incorrectly, determines how the testee moves through the structured item pool.

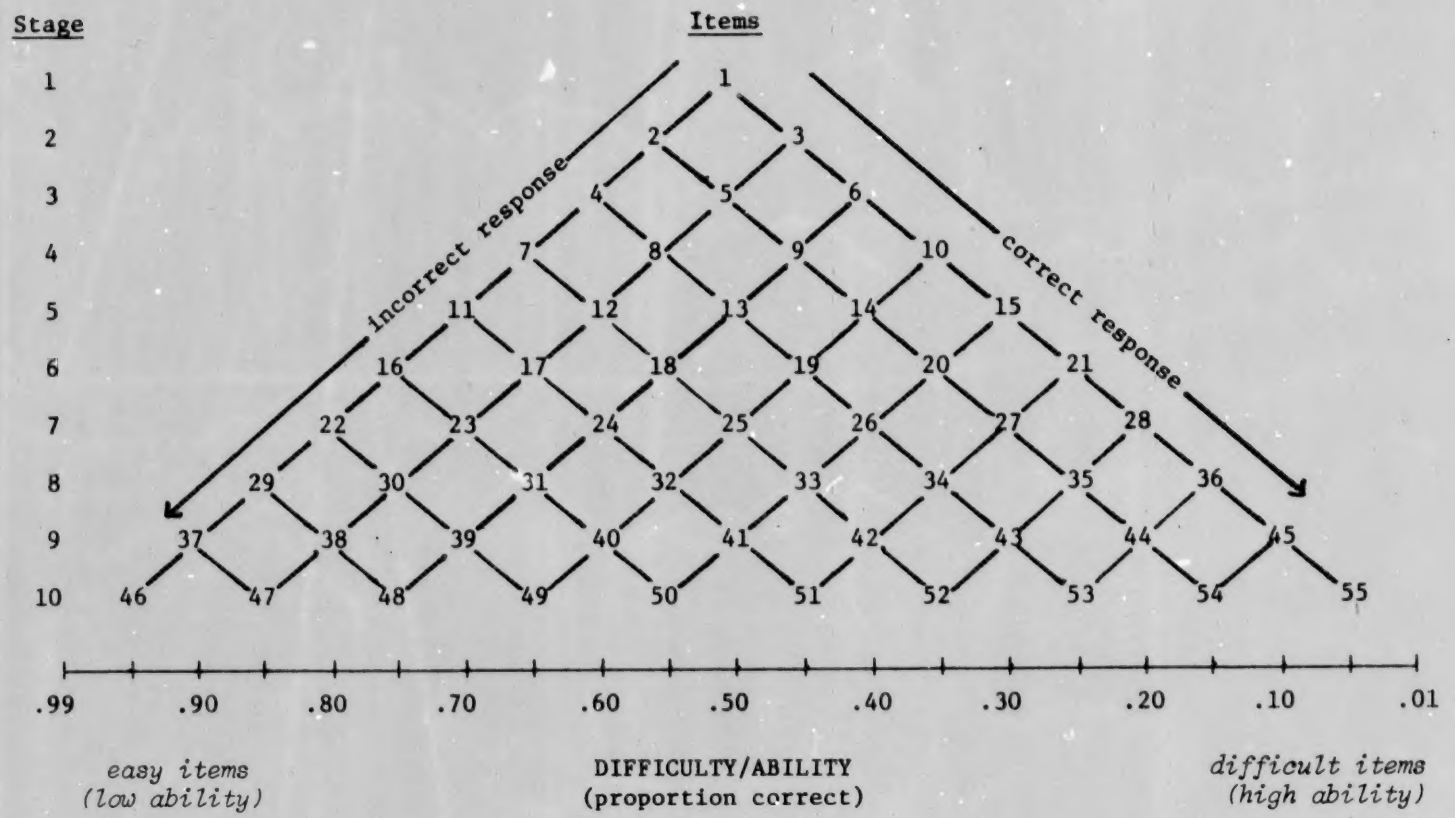
Pyramidal Models

The pyramidal, or "tree-structure," models were the first adaptive testing models proposed and have generated the most research to date. Research on pyramidal models (Weiss & Betz, 1973) was initiated by Krathwohl and Huyser (1956), continued by Bayroff (Bayroff, Thomas & Anderson, 1960; Bayroff & Seeley, 1967), Paterson (1962), Hansen, (1969) and Lord (1970, 1971a), among others, with the most recent contributions made by Mussio (1973) and Larkin and Weiss (1974). Many variations of the pyramidal models have been proposed and these can be differentiated into those using constant step sizes, variable (decreasing) step sizes, truncated pyramids, multiple-item pyramids, and pyramids using differential response option branching.

Constant step size pyramids. Figure 4 shows the tree-like item structure of a 10-stage pyramidal test with constant step size. Constant step size pyramids require that the number of items available at each stage be equal to the rank of the stage. Thus, at stage 1 there is one item available, at stage 5 there are five items, and at stage 10 there are ten items available. A 10-stage pyramid structured in this way requires 55 items.

The base line of Figure 4 shows the difficulties associated with the items in the pyramid. Item difficulties range from p=.95 (i.e., 95% of the norming group answered the item correctly) to p=.05 (5% answered correctly). The vertical columns of items

Figure 4
A Pyramidal Item Structure With Constant Step Size



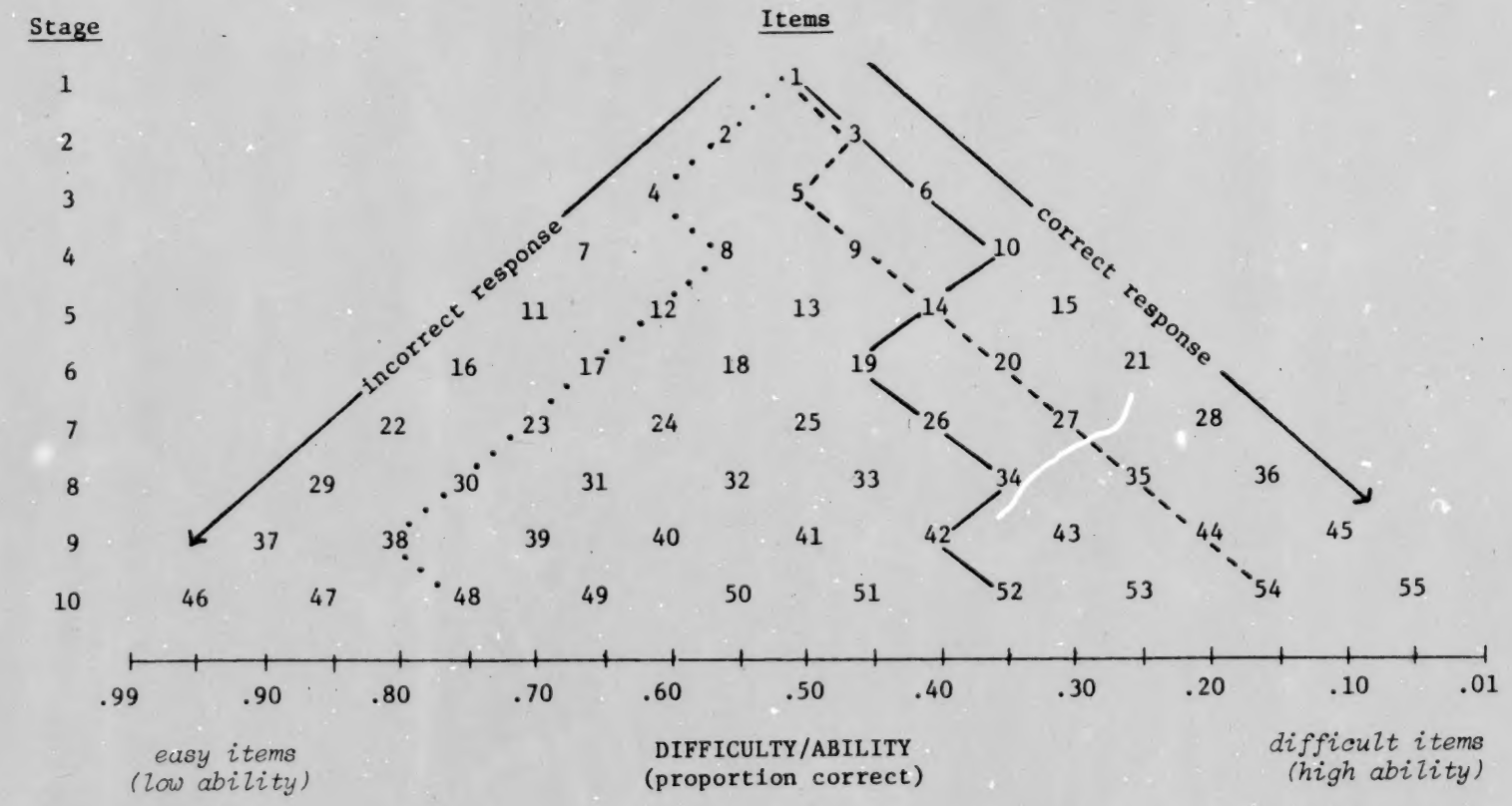
at each difficulty level in Figure 4 indicate items of similar difficulty. In Figure 4, only one item (no. 46) has a difficulty of .95, one item has a difficulty of .90 (item 37), and five items have difficulties of .50 (items 1, 5, 13, 25, 41).

Movement through the pyramidal structure begins for all testees at stage 1 (item 1 in Figure 4). The testee's response to this item is scored as correct or incorrect, the branching rule is consulted, and the appropriate stage 2 item is then administered. A typical branching rule is "up-one/down-one." Using this branching rule, following a correct response to an item the testee receives an item one increment higher in difficulty. Following an incorrect response he is branched to an item one increment lower in difficulty, i.e., to a slightly easier item.

For example, in Figure 4 the items at successive stages differ in difficulty by a constant step size of p=.05. Using an up-one/down-one branching rule, an incorrect reponse to item 1 (difficulty of p=.50) leads to item 2, which has a difficulty of p=.55. An incorrect response to item 2 leads the testee to item 4, with a difficulty of .60, while a correct response to item 2 leads to item 5, another item of the same difficulty as item 1. The testee continues to branch through the item pool, having each item response evaluated at each stage, and receiving a slightly more difficult item following each correct answer and a slightly easier item following an incorrect answer. The testee answers only one item at each stage, and testing continues until each testee answers an item at stage 10, or the nth stage if the pyramid consists of other than 10 stages. Thus, each individual can follow a number of paths from the stage 1 item to an item at stage n, receiving only one item at each stage.

Figure 5 shows three illustrative paths through the 10-stage pyramid shown in Figure 4 (using a constant step size of p=.05, and an up-one/down-one stepping rule). All three paths begin at item 1. The path traced by the solid line shows a testee of slightly above average ability. His response to the first item (item 1), an item of .50 difficulty, was correct; he was thus branched to item 3, a slightly more difficult item (p=.45) which he also answered correctly. The stage 3 item was item 6 with a difficulty of p=.40, which was also answered correctly, resulting in the administration of item 10. Item 10 (p=.35) was the first item answered incorrectly. The solid line thus branches to the left leading to item 14 (p=.40) which was also answered incorrectly. An incorrect response to item 14 led to item 19, an item of .45 difficulty. This item was answered correctly, as was item 26, leading the testee back to the more difficult items. At item 34, which was answered incorrectly, the testee began to alternate between correct and incorrect responses to items of .35 and .40 difficulty (items 34, 42, 52). Finally, the testee reached item 52, the stage 10 item of .35 difficulty.

Figure 5
Illustrative Paths through a Constant Step Size Pyramid



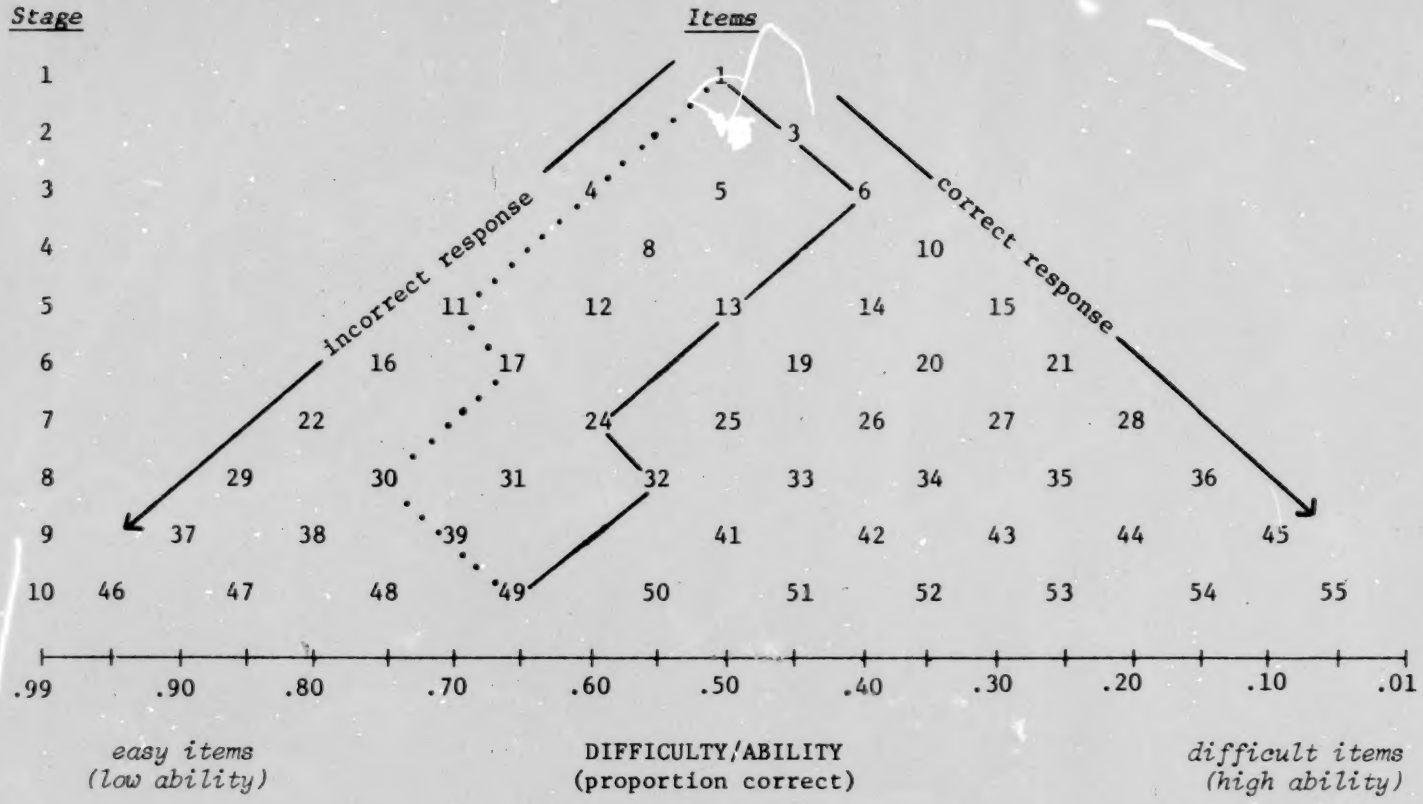
The dashed line in Figure 5 shows a different path through the pyramidal structure, for a testee of slightly higher ability than that depicted by the solid line. With the exception of an incorrect response to item 3, the test record traced by the dashed line shows all correct responses, leading to items of higher and higher difficulty. The final item reached in the dashed path is item 54, one of two items with difficulties of .15. The individual traced by the dotted line, on the other hand, gave only two correct answers--to items 4 and 38. The difficulty of the last item reached by the dotted path was .75, an easy item which 75% of the norm group answered correctly.

All of the examples in Figure 5 used an up-one/down-one branching rule for selecting the next item to be administered. The up-one/down-one rule uses an equal "offset"--the number of steps branched to more difficult items is the same as the number of steps branched to easier items. With multiple-choice items on which guessing is possible, it might be desirable to slow somewhat the branching to more difficult items so that unnecessarily difficult items are not presented to a testee following a chance success. This involves using an unequal "offset." An unequal offset results in branching differently in one direction than the other. To reduce the effects of guessing, the tester might wish to use an up-one/down-two branching rule, which implies an unequal offset.

Figure 6 is an example of two test records using an up-one/down-two branching rule where each correct response leads to an item .05 higher in difficulty and each incorrect response leads to an item .10 lower in difficulty. The test record traced by the solid line shows correct responses to item 1 and 3, leading by steps of .05 to item 6. Item 6 (p=.40) was incorrectly answered. Thus, following the branching rule, the next item was of .50 difficulty--item 13. That item was answered incorrectly leading to item 24 (p=.60). Item 24 was answered correctly so item 32 was administered (p=.55). Finally, item 32 (p=.55) was answered incorrectly, and the last item administered was item 49 (p=.65). The path shown by the dotted line includes only three correct responses, to items 11, 30, and 39. Each correct response resulted in the administration of an item .05 higher in difficulty, while the incorrect responses resulted in items .10 lower in difficulty.

When using the standard pyramid structure with other than an equal offset, the number of items administered to an individual will usually not equal the number of stages. Thus, the two paths shown in Figure 6 show that those testees each completed only 7 items in the 10-stage pyramid. The number of items to be administered to any testee, when the offset is unequal, will vary as a function of the number of correct responses. With an up-one/down-two branching rule in a 10-stage pyramid, ten items will be administered if all answers are correct, while only five items will be administered if all answers are incorrect.

Figure 6
Paths through a Constant Step Size Pyramid with Up-one/Down-two Branching Rule

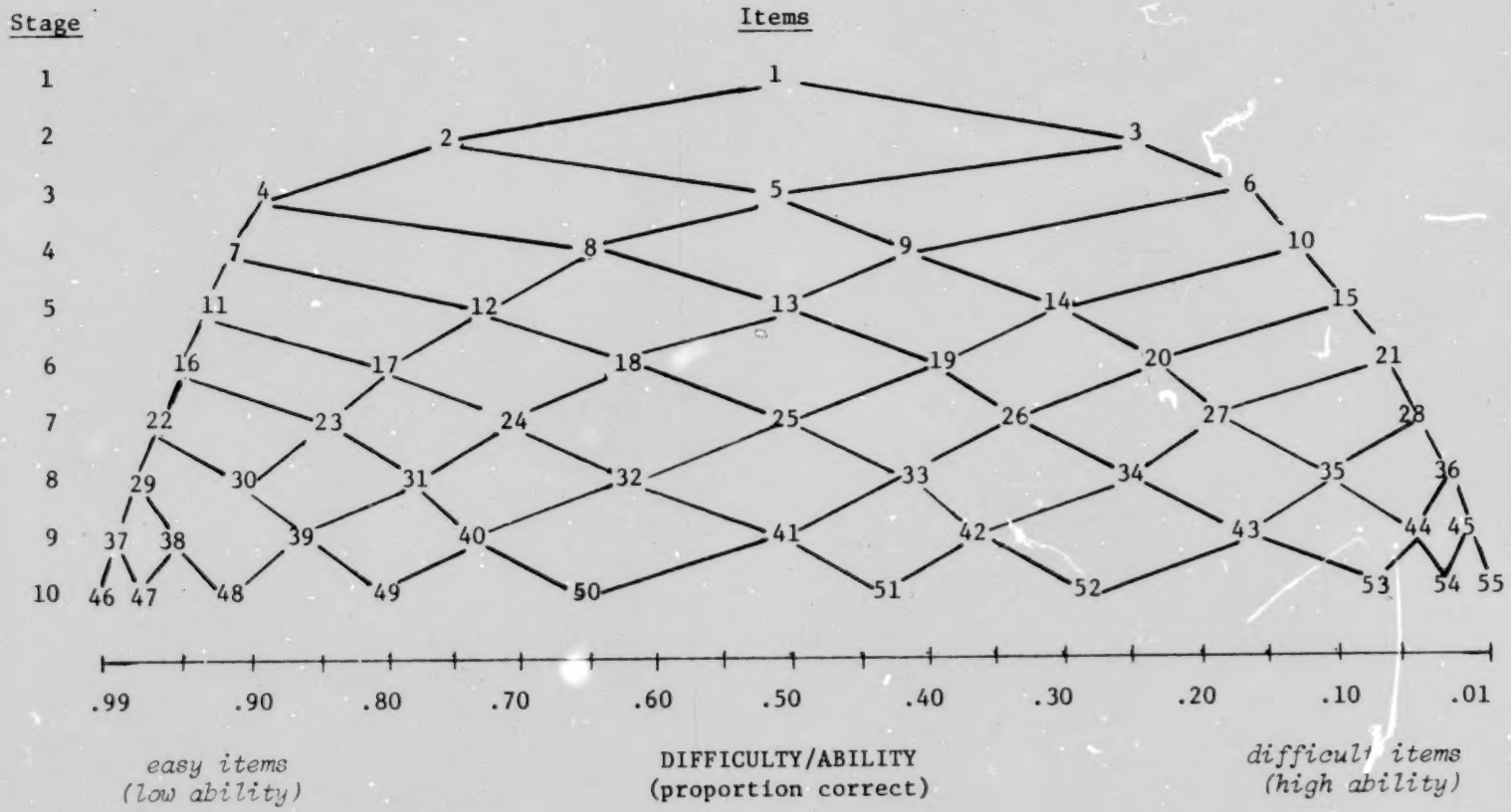


Decreasing step size pyramids. Adaptive testing should permit the tester to converge upon a region of the item pool which is most appropriate for a given testee. Thus, an adaptive test should permit the tester to identify, in as efficient a way as possible, the test items which are at the individual's level of ability, as indicated by the fact that he answers about half of those items correctly and half incorrectly (Lord, 1970). Paterson (1962) and Lord (1971a), among others, have suggested that constant step size pyramids are an inefficient way to locate this region of the item pool quickly. Instead, they suggest that large step sizes be used in the early stages of testing, and that step sizes should decrease at the later stages of testing in order to converge more precisely on an individual's ability level.

Paterson (1962) described an item structure for a shrinking step size fixed branching pyramidal test. Figure 7 shows such an item structure. The first item administered is, as is typical of pyramidal tests, an item of median (p=.50) difficulty. Stage 2 items (items 2 and 3) are placed midway between the stage 1 item and the extremes of the difficulty distribution, at p=.75 and p=.25, respectively. Thus, the step size in moving from stage 1 to stage 2 is .25; this contrasts with the step size of p=.05 used in the previous examples. The effect of the larger step size is to move the testee to the center of the upper and lower halves of the ability distribution as a first estimate of his ability level. The three stage 3 items are located midway between the stage 2 items, or between the the stage 2 items and the extremes of the ability distribution. Thus, the difficulty of item 4 would be approximately .875; that of item 5 would be p=.50; and that of item 6 would be p=.125. The step size for moving from stage 2 to stage 3 would be .125 for the extreme items (items 4 and 6), or half the step size for stages 1 to 2, and .25 for moving from items 2 or 3 to item 5. Step sizes for the remaining items are computed in an analogous way. For items at the upper and lower ends of a stage (e.g., items 7 and 10), the distance between the extreme item at the stage immediately above it and the highest (or lowest) difficulty possible is divided in half to obtain the difficulty of the item in question. For items between the highest and lowest at a stage (e.g., items 8 and 9) the appropriate difficulty is halfway between the difficulties of the two items above it at the preceding stage.

The effect of structuring a pyramid in this way is to obtain step sizes which are progressively smaller from one stage to the next. Because of the upper and lower limits of the difficulty distribution, step size decreases more quickly for items near the extremes of the difficulty distribution than it does for items near the center of the difficulty distribution. The pyramidal structure in Figure 7 has the effect of concentrating more test items at the extremes of the ability/difficulty distribution as compared to the fixed step size procedure in which more items are concentrated near the center of the ability distribution (e.g., see Figure 4). An important feature of the pyramidal

Figure 7
Item Structure for a Decreasing Step Size Pyramidal Test



structure of Figure 7 is that is is likely to provide scores of more equal precision at all ability levels, without requiring any more items than does the fixed step size procedure.

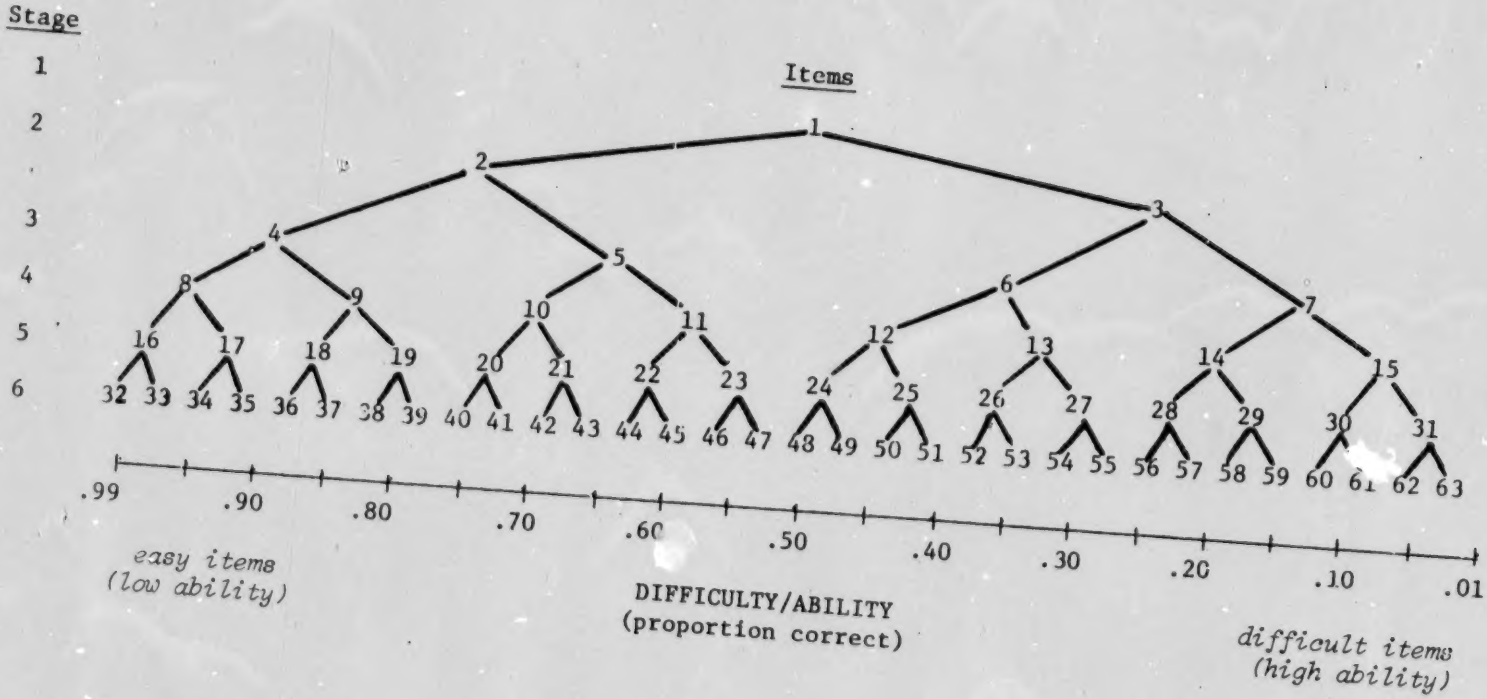
Lord (1971a) has proposed another shrinking step size procedure designed to converge more precisely on an individual's ability level. Lord's method is based on a "Robbins-Monro" convergence procedure. A schematic representation of one such procedure is shown in Figure 8. The Robbins-Monro pyramidal structure shown in Figure 8 is for a 6-stage test, i.e., each testee will answer only 6 items. The stage 1 item has difficulty p=.50; stage 2 items are of difficulty .75 and .25 respectively, resulting in a step size of .25. The step size for branching from stage 2 to stage 3 is half that for movement from stage 1 to stage 2, or p=.125. Thus, item 4 has difficulty .875; item 5, p=.625; item 6, p=.375; and item 7, p=.125. The step size for branching from stage 3 to stage 4 is then half the original step size, or p=.0625. Thus, the items at stage 4 have difficulties .0625 above or below those of the stage 3 items. In the example shown, the halving of step sizes continues so that the difficulty increment of items at any stage is half that of the previous stage. Other methods for reducing the step size at each stage are also possible (e.g., Lord, 1971a). Regardless of how the step size is reduced, the Robbins-Monro procedure requires a doubling of items at each stage over the number available at each earlier stage. Two items are required at stage 2, 4 at stage 3, 8 at stage 4, and so on. In general, as is shown in Figure 8, 2^n-1 items are required at the nth stage of the Robbins-Monro pyramid; thus, stage 6 alone requires 32 items. The total number of items required is 2^n-1, or $2^6-1 = 63$ items for the 6-stage Robbins-Monro pyramid in Figure 8.

The Robbins-Monro procedure promises rapid convergence on ability at all levels of ability, as compared to the fixed-branching procedures which promise more accurate measurement of abilities near the mean, and Paterson's non-Robbins-Monro shrinking step procedure which will probably measure more accurately at the extremes. However, as Lord (1971a) has pointed out, the Robbins-Monro procedures require prohibitively large numbers of items in the pyramidal structure for even moderate-sized pyramids. The 6-stage Robbins-Monro structure of Figure 8 requires 63 items, while the 10-stage constant step size pyramid of Figure 4 requires only 55 items.

Another limitation of the Robbins-Monro structure is its susceptibility to chance successes. All pyramidal structures used with multiple-choice items are susceptible to branching errors resulting from chance successes due to guessing. However, the capability of the Robbins-Monro procedure to recover from these branching errors is substantially less than that of the other pyramidal structures, particularly when guessing results in chance successes early in testing. For example, consider the case of the low-ability individual (say of ability corresponding to p=.80) who made a lucky guess on item 1 of a Robbins-

Figure 8

Pyramidal Item Structure for a Six-stage Robbins-Monro Shrinking Step Size Pyramidal Test



Monro procedure. Because of that lucky guess, his correct response to that item leads him to item 3 of difficulty p=.25. Since item 3 is obviously too difficult, he answers incorrectly and is branched downward to item 6 (p=.375) which is also too difficult. Assuming that he had no further lucky guesses and answered all subsequent items incorrectly, the reducing step size procedure shown would take an infinite number of items to reach a difficulty level of p=.50, still well above the testee's true difficulty level of p=.80. Other ways of reducing the step size under the Robbins-Monro approach will result in very large numbers of stages necessary for complete recovery from early chance successes. Thus, the Robbins-Monro procedure appears to be limited to free-response items where there is virtually no chance of correct responses occurring as a result of guessing.

The decreasing step-size pyramid in Figure 7, however, is much less susceptible to the same guessing effects. Assuming the same lucky guess on item 1 (and the same unlucky guesses on all subsequent items that are too difficult) the pyramid of Figure 7 will return the testee to his approximate ability level near p=.80 by stage 5, via items 5, 8 and 12. On the other hand, the fixed step size pyramid shown in Figure 4 would require seven items to reach a terminal difficulty level of p=.80 after the same lucky guess. Thus, it appears that if multiple-choice items are to be used in pyramidal item structures so that chance successes from guessing are likely to occur, the non-Robbins-Monro decreasing step size pyramidal structure has the most logical appeal, while the Robbins-Monro procedure should not be considered.

It should be noted, additionally, that Figures 7 and 8 represent only two of many decreasing step size procedures possible. It is also possible to mix fixed and decreasing step size procedures at various stages of pyramidal testing. This would maximize the degree to which the tester might achieve accuracy of measurement at various points of the ability distribution and economy of items in terms of rapid convergence on ability levels.

Truncated pyramids. While Robbins-Monro procedures require very large numbers of items to form a pyramid, non-Robbins-Monro decreasing step size pyramids and fixed step size pyramids also make fairly heavy demands on an item pool. The 10-stage pyramids of Figures 4 and 7 each require 55 items. In general, a pyramidal item structure requires $n(n+1)/2$ items where n is the number of stages. Thus, a 20-stage pyramid would require 210 items in its structure. If pyramidal tests are to approximate the length of conventional tests (although such long pyramids are not really necessary to achieve measurement efficiency; see Weiss & Betz, 1973) very large item pools will be needed.

Mussio (1973) has proposed a method of reducing pyramidal item pool requirements. His proposal is based on a Markov chain

stochastic model with reflecting or retaining barriers. In essence, Mussio proposed truncating the tails of the pyramid (i.e., eliminating items at the extreme levels of difficulty). Once the truncation occurs, two kinds of branching are possible, based on whether a reflecting barrier or retaining barrier is used.

Figure 9 shows the truncated pyramid structure with reflecting and retaining barriers, for a 10-stage pyramidal test. For the reflecting barrier, items are concentrated between difficulty levels of .65 to .35 only; all items above and below these difficulty levels are eliminated from the structure, effecting a savings of 24 items for the 10-stage pyramid. Branching occurs in the usual way, with correct answers leading to more difficult items and incorrect answers leading to less difficult items through the first four stages. When a testee gives an incorrect answer to item 7, or a correct answer to item 10, the reflecting barrier takes effect.

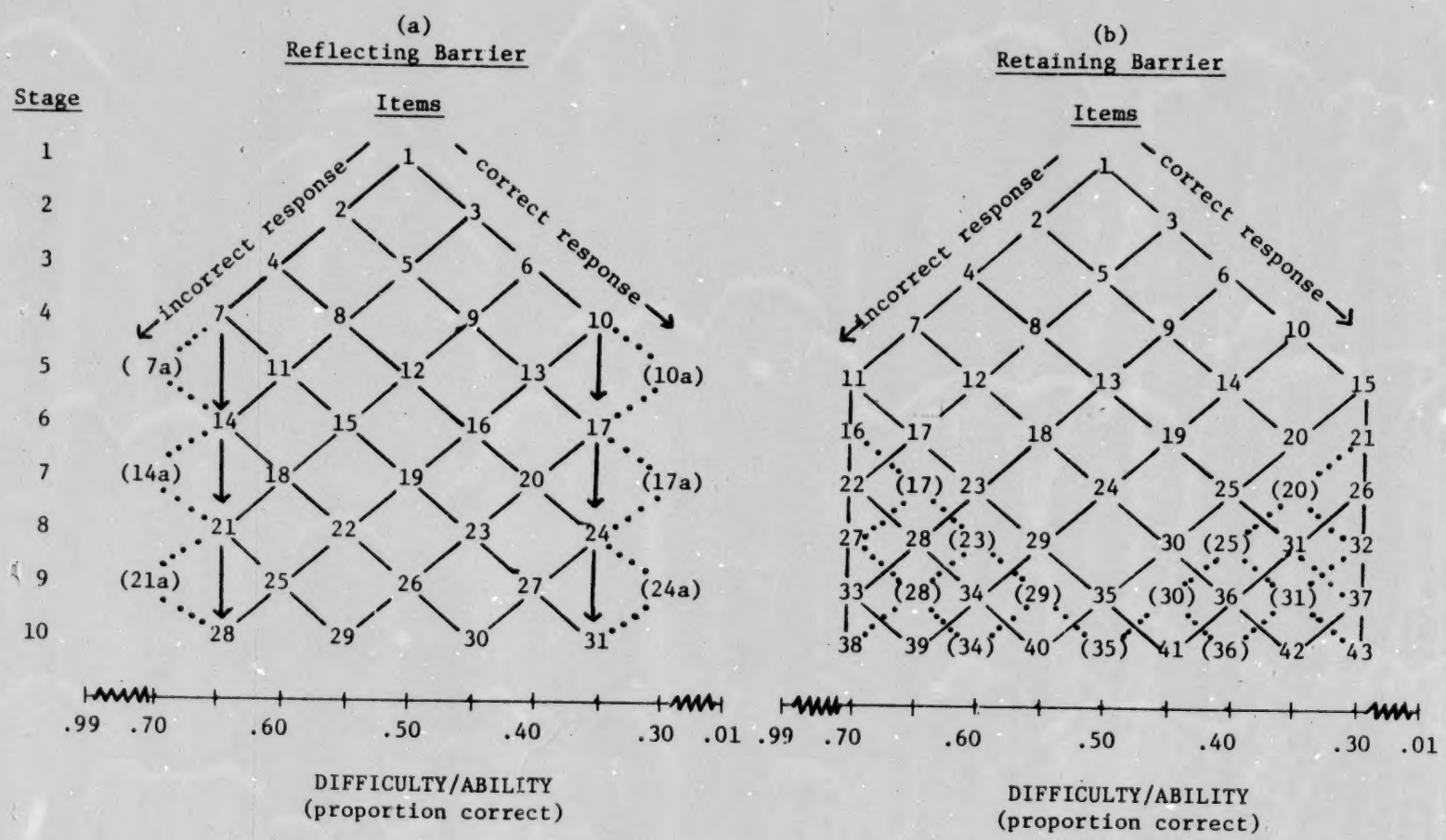
Items 7a, 14a and 21a represent the items that would be administered following an incorrect response to items 7, 14, or 21. However, because the pyramid is truncated using a reflecting barrier, either a correct or incorrect reponse to these items leads the testee to the same item (e.g., item 7a branches only to item 14). As a result, these items need not be administered.

Thus, using a reflecting barrier, an incorrect response to item 7 will branch the testee to item 14, while a correct response to 7 will branch to item 11. A major characteristic of the reflecting barrier, therefore, is that items at the barrier (the point of truncation of the pyramid) are only hypothetical items that provide no differential branching. They can, consequently, be eliminated from the item structure which results in fewer than n items administered to some testees in an n-stage test. It also results in a savings in the number of items required for the pyramidal structure, since only 31 items are required for a 10-stage pyramid. At the same time, however, all items at or beyond two levels of difficulty (i.e., those greater than .65 and less than .35) are eliminated from the test, thereby reducing the effective range of measurement.

The retaining barrier approach, on the other hand, retains the items at the two levels of difficulty at which the pyramid is truncated. This is accomplished by adding additional items at the barriers with difficulties the same as those already available. The retaining barrier branching diagram in Figure 9b shows three additional items available at p=.70 (items 16, 27 and 38) and three at p=.30 (items 21, 32 and 43).

Branching in the retaining barrier approach occurs in the usual way until the barriers are reached. The testee who answers items at the first four stages incorrectly will be branched through easier items from item 1 through items 2, 4 and 7 to item 11. A correct response to item 11 will branch to item 17. An incorrect to item 11, however, leads to item 16, an item of the same

Figure 9
Truncated Pyramidal Item Structures with Reflecting and Retaining Barriers



difficulty as item 11. As long as the testee answers incorrectly he will continue to receive items of the same difficulty.

When a testee gives a correct answer to one of the additional items at the retaining barriers (e.g., items 16, 27 or 33 at the p=.70 difficulty level in Figure 9b) he receives a more difficult item. To accomplish this, the branching network is, in effect, shifted down one stage so that a correct response to item 16 at stage 6 leads to item 17 at stage 7, as shown by the dotted lines in Figure 9b. Subsequent branching using the retaining barrier approach would follow the path shown by the dotted lines, using item numbers as shown in the parentheses. As a result of this procedure (and an analogous procedure at the upper retaining barrier), each testee completing a pyramidal test using a retaining barrier will complete the same number of items.

Mussio (1973) provides formulas for determining the number of items required for the reflecting barrier and retaining barrier approaches. In comparison with the reflecting barrier, which required 31 items, the 10-stage retaining barrier in Figure 9b requires 43 items. The reflecting barrier pyramid in Figure 9a has items available at only seven levels of difficulty and the retaining barrier has items at nine levels of difficulty; a complete pyramidal structure (e.g., Figure 4) includes 55 items at 19 levels of difficulty.

As proposed by Mussio, reflecting and retaining barrier pyramidal item structures can be used with either constant or variable step sizes. The major advantage of the methods is in item economy. For a 60-stage pyramid either of the truncated pyramids requires less that 25 percent of the number of items used in a full pyramidal structure, if the truncated pyramids are confined to items at eleven levels of difficulty. The truncated pyramids would also appear to be less susceptible to guessing effects since the testee can return more quickly to the main part of the branching structure than would be possible after several chance successes in a standard pyramidal test. The major deficiency of the reflecting and retaining barrier approaches is that they concentrate measurement around the mean of the distribution of item difficulties, reducing the capability of the test to make discriminations among testees whose abilities fall near the upper and lower extremes of the ability distribution.

Multiple-item models. Several writers (e.g., Krathwohl & Huyser, 1956; Linn, et al., 1969) have proposed or studied pyramidal branching models with more than one item per stage in an attempt to improve the reliability of branching decisions and/or to reduce the number of stages in the multi-stage model. In general, in these models all items at one stage are scored before the items to be administered at the next stage are selected. Figure 10 shows an example of a "three items per stage" multi-stage pyramidal testing model.

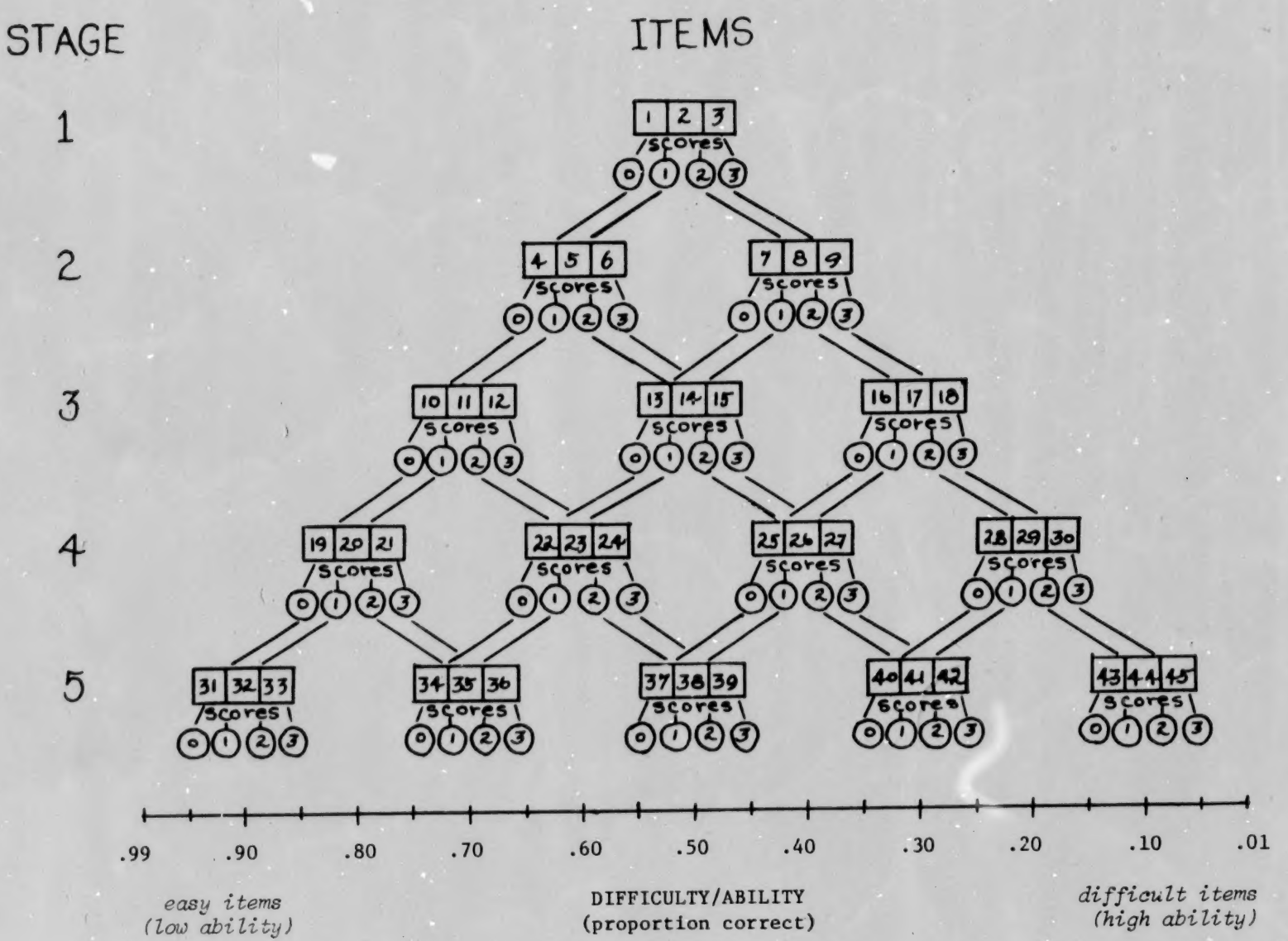
The first three items in the pyramid shown in Figure 10 are items with difficulties between p=.55 and p=.45. The three items are administered to the testee and his score on the three stage 1 items is determined. Testees with scores of 0 or 1 are branched to items 4,5 and 6 at stage 2, which are easier items with difficulties in the range p=.56 to .65; testees with scores of 2 or 3 on the stage 1 items are branched to items 7, 8 and 9 at stage 2, which are more difficult items in the difficulty range p=.44 to .35. The procedure is the same at each successive stage in the testing procedure--three items are administered at each stage, the three items are scored, and branching to the next stage is based on the number correct at the previous stage. At the final stage (stage 5 in Figure 10) scores on the items are then used to obtain a wider range of final scores on the pyramidal test than would be available from a pyramidal test of an equal number of stages (but not an equal number of items). If the number of items answered correctly at the final stage and the difficulty of the items at that stage are both considered, the 5-stage pyramidal structure in Figure 10 results in twenty possible scores. A typical 5-stage pyramid would result in a maximum of only ten scores (see below). However, the multiple-item pyramid requires that each testee complete fifteen items while the pyramid with one item per stage requires only five responses from the testee.

There are obviously many possible variations of the multiple-item pyramidal structure. Such pyramids can use constant or variable step sizes, and equal or unequal offsets. Multiple-item pyramids could be constructed with varying numbers of items per stage. To make more gross discriminations at the initial stages of testing and finer discriminations at the later stages, such pyramids could use smaller numbers of items per stage at the earlier stages and larger numbers per stage at later stages. The distribution of item difficulties within a stage could vary; thus, each stage could be a short "peaked" test or it could be a narrowly distributed rectangular test. Branching decisions based on the scores at each stage could also be varied. In this approach, the step size could be a function of the number of items answered correctly at each stage. Thus, these variations of the multiple-item pyramid reflect their nature as hybrids combining elements of two-stage models and the pyramidal models.

The major advantage of this method, however, appears to be that administering more than one item at each stage will lead to branching decisions which are less influenced by chance successes. The resulting final scores, then, which are based on more items, should be more reliable, thus improving accuracy of measurement.

Differential response option branching. The objective of this procedure is to utilize all of the information in a testee's response to a test item by branching to different items at the

Figure 10
A Three-items-per-stage Pyramidal Test Structure



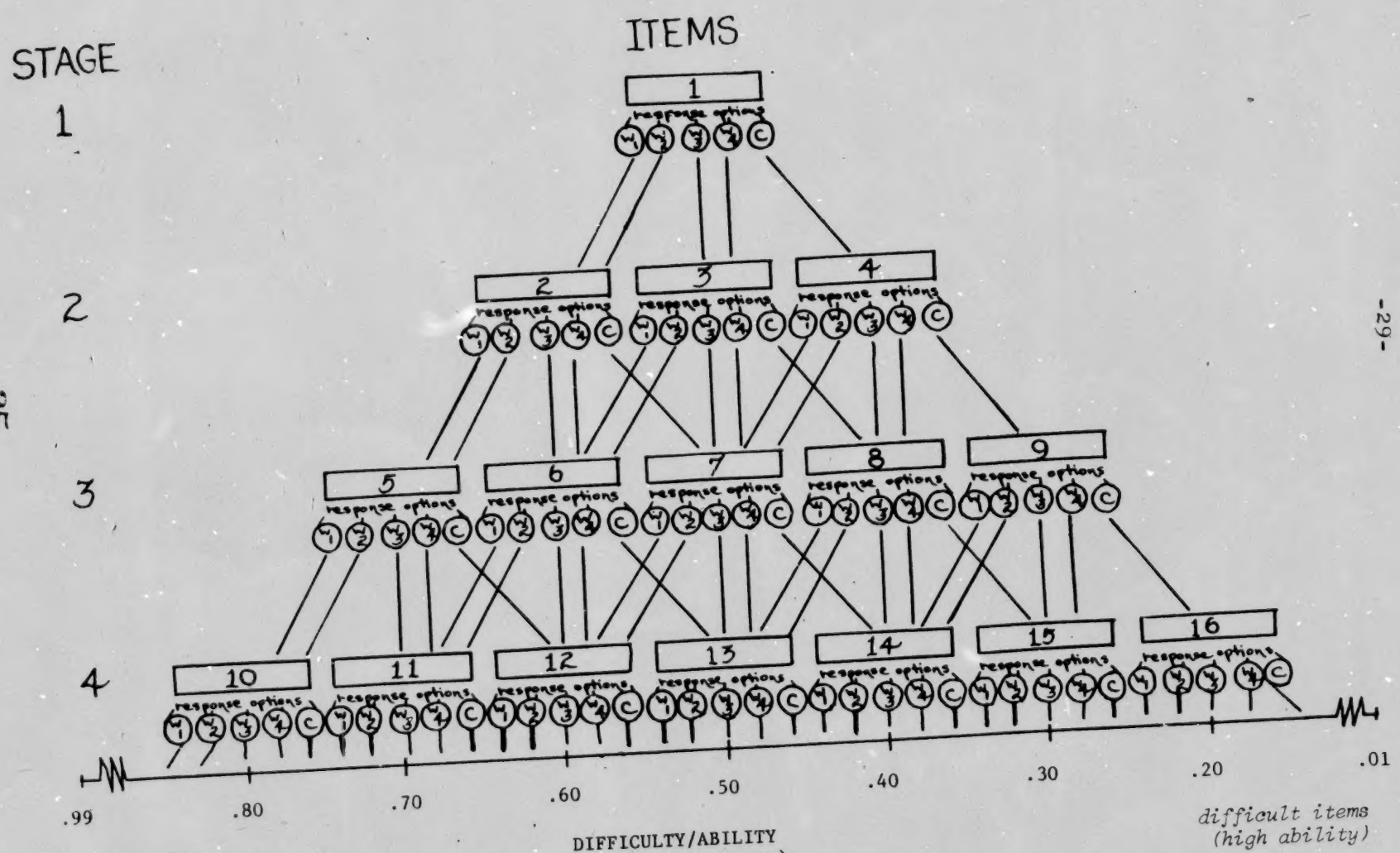
next stage based on the degree of "wrongness" of an incorrect response. This pyramidal branching model is designed primarily for use with multiple-choice test items. It can, however, be applied to free-response items where there is a relatively limited number of answers that can be graded in terms of difficulty. The use of differential response option branching was originally suggested by Bayroff (Bayroff & Seeley, 1967). Bayroff used the approach only on the first stage of an otherwise standard 5-stage pyramidal item structure.

Figure 11 shows one way of operationalizing the differential response option branching approach. There is one item available for administration at stage 1, 3 items at stage 2, 5 at stage 3, and so on, with the number of items increasing by two at each stage. The items depicted in Figure 11 are 5-alternative multiple-choice items, with one correct reponse (C) and four "incorrect" responses (W_1, W_2, W_3 and W_4). For purposes of illustration, response alternatives W_1 through W_4 for each item are ordered in terms of "correctness," with W_1 being the least correct and W_4 being the incorrect alternative which is most nearly correct. "Correctness" is a function of proportion of persons in the norming group who choose each distractor, or the normal ogive difficulty values for each distractor (Lord & Novick, 1968). Correctness can also be thought of as reflecting the average ability level of individuals choosing each alternative.

At stage 1 the testee is presented with item 1. His response is recorded and is categorized as either a correct response (C) or one of the four incorrect responses W_1, W_2, W_3 or W_4. If the response is C the testee is branched to item 4 at stage 2, a more difficult item. If the response is W_1 or W_2, the two least correct alternatives, the testee is branched to an easier item, item 2. For responses W_3 and W_4, which are alternatives of higher difficulty than W_1 and W_2, yet are not correct, the testee is routed to item 3, an item of the same difficulty as item 1. The logic of the procedure is that W_3 and W_4 are more frequently chosen by individuals of average ability, rather than those of higher or lower ability, so that branching to items at a higher or lower level of difficulty is not really appropriate; hence branching occurs to an item of the same level of difficulty.

Similar branching decisions are made following the responses to items at each successive stage. A choice of the correct response leads to an item of higher difficulty, a choice of either of the two most incorrect response alternatives leads to an item of lower difficulty, while choices of the intermediate difficulty alternatives lead to an item of the same difficulty. The example in Figure 11 shows a major advantage of differential option branching. If the "correctness" of the answer chosen to the item at the final stage is used as the testee's score on the test, a four-stage pyramid of this type

Figure 11
A Pyramidal Item Structure Using Differential Response Option Branching



results in thirty-five possible different scores, while a typical four-stage pyramid would yield only eight possible score values. The differential option branching pyramid accomplishes this by utilizing all the information available in the testee's response, rather than combining all incorrect answers into one score category.

The example of Figure 11 is only one way of operationalizing differential response option branching. Obviously, complete information utilization would require differential branching for each available response alternative. Thus, rather than branching to the same next item for both W_1 and W_2, this procedure would require that W_1 responses branch to a less difficult next stage item than W_2 responses. The difficulty of the next stage item would be approximately the same as the scaled difficulty of the stage 1 alternative chosen. This would require five possible branches for each 5-response-choice multiple-choice item, rather than the three shown in Figure 11. It would also require that more easy items be available in the branching network. As a result, the pyramid would become asymmetric, with more items available on the left side to allow for the greater number of branches resulting from the various response options. Such an item pool would be considerably larger than the pool required for a simple correct-incorrect branching procedure. However, the potential gains in accuracy of measurement, with a constant number of items administered to each testee, would have to be weighed against the item pool requirements. It is obvious, however, that differential response option branching is a fertile area for psychometric research.

Scoring. A number of scoring schemes have been proposed which can be used in all variations of the pyramidal model. These scoring methods are, in large part, based on the difficulties of the items answered by the testee. Thus, it is assumed that the pyramidal item pool is unidimensional and that the difficulty scale and the ability scale can be expressed in the same terms.

One method of scoring pyramidal tests uses as the testee's score the difficulty of the most difficult item answered correctly. This scoring method assumes the "maximum performance" conception of ability testing. If guessing is possible, however, this scoring method might lead to unreliable scores because of chance successes. A related scoring method determines score as difficulty of the final item. Since the pyramidal test should, for most individuals, converge upon a difficulty level appropriate for each individual, the difficulty of the final item reached should reflect the individual's ability level. Where the number of stages in the test is small, however, the resulting number of unique scores will be quite small. Figure 4, for example, shows a 10-stage pyramid which has 10 terminal items, resulting in only 10 possible scores by this method of scoring.

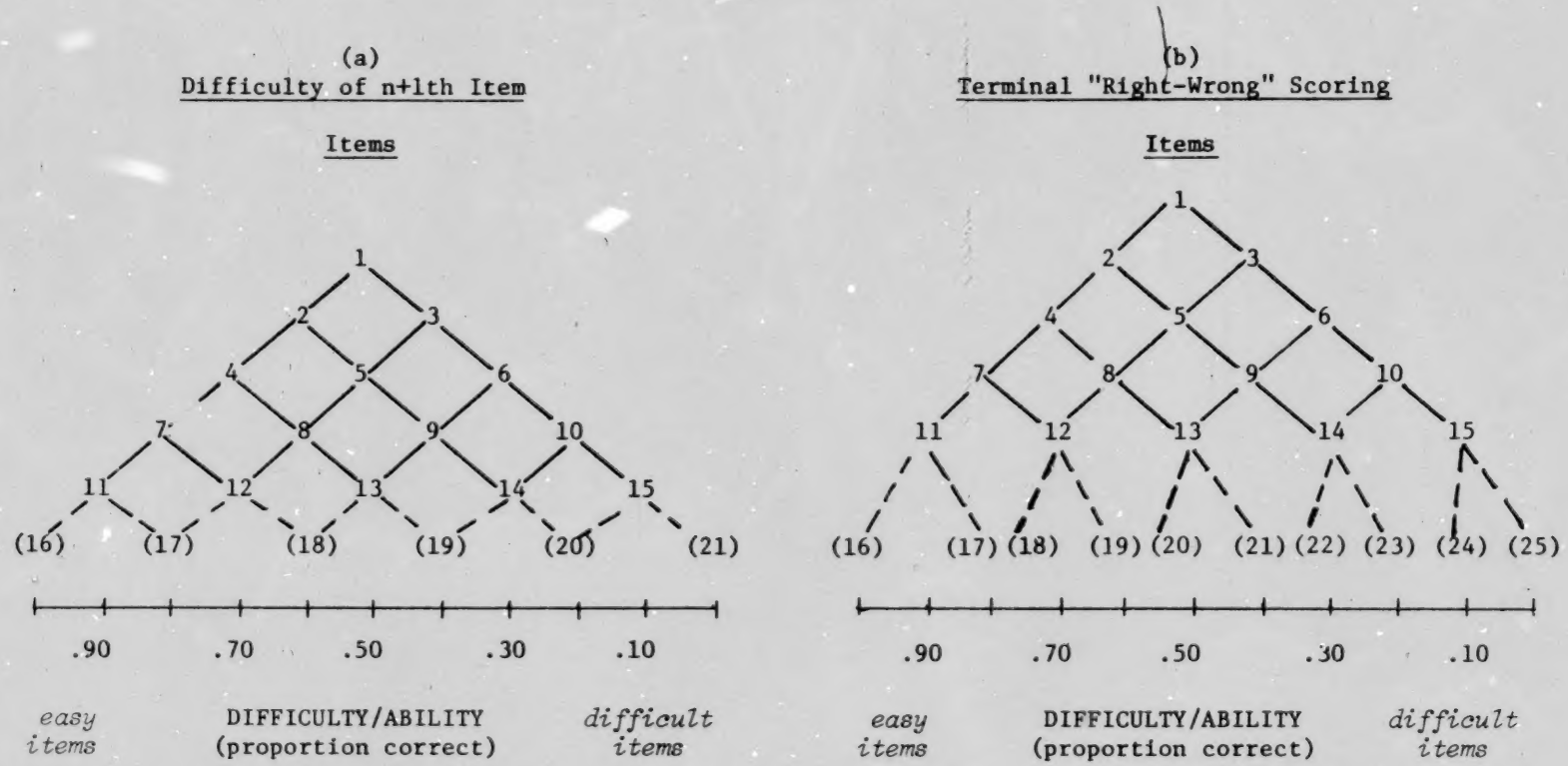
Such a restricted number of scores could result in lower correlations of pyramidal tests with other variables, e.g., validity criteria, thus reducing the practical utility of this testing strategy.

The final difficulty score does not take account of the testee's response to the last item administered. Thus, a testee who answers the terminal item incorrectly would get the same score as the testee who answers the final item correctly. Using the difficulty of the "(n+1)th" item has been suggested as a way of differentiating between those who answer the final (nth) item correctly and those who answer it incorrectly. In essence, this method assumes the existence of another stage of hypothetical test items following the nth, or last stage of items actually administered. Figure 12a illustrates this for a 5-stage pyramid.

As Figure 12a shows, for a 5-stage pyramid there will be five terminal items (items 11 through 15). Adding an (n+1)th stage of hypothetical items (items 16 through 21) takes account of the testee's response to the terminal item. For example, item 12 might have a difficulty of p=.70, which would be the testee's score under the "terminal difficulty" scoring method. Using difficulty of the (n+1)th item, the testee who answers item 12 incorrectly would be "branched" to the hypothetical item 17, with a difficulty of .80; .80 would then be his final score. The testee who answers item 12 correctly would be assigned a final score of .60, the difficulty of hypothetical item 18. While scoring a pyramidal test using difficulty of the (n+1)th item makes use of the testee's response to the last item, it does little to increase the range of scores resulting from pyramids with small numbers of stages.

The "terminal-right-wrong" scoring method, illustrated in Figure 12b, makes use of the testee's response to the last item and, at the same time, increases the number of possible scores derivable from the pyramidal test. This method, in contrast to the (n+1)th scoring method, is based on the assumption that a correct response to an easier item (e.g., item 12) is not indicative of as high ability as an incorrect response to a more difficult item (e.g., item 13). This scoring method also uses hypothetical items following the last stage, but there are two such items for each terminal item. Thus, as Figure 12b shows, item 12 branches to the two hypothetical items numbered 18 and 19. The difficulty of item 18 is somewhat lower than that of item 12 (e.g., .75 and .70 respectively) while item 19 (p=.65) is slightly more difficlut than item 12. In contrast to the (n+1)th scoring method, no hypothetical item can be reached from two different stage n items. Thus the difficulties of hypothetical item 19 (p=.65) and hypothetical item 20 (p=.55) are not the same. As a result, different scores are assigned to testees whose responses fall into the two categories. The effect is to double the number of possible scores resulting

Figure 12
Pyramidal Scoring Based on Two Methods of Using Final Item Response Data



from the pyramidal test.

Research with pyramidal models also has used average difficulty scores. As used by Lord (1970), this method ignores the difficulty of the first item administered, since it is the same for all individuals, but includes the difficulty of the (n+1)th item. In computing the average difficulty score, Lord includes all items encountered by the testee, regardless of whether he answered the item correctly or incorrectly. This method, therefore, represents the average item difficulty of the testee's complete path through the pyramid. Average difficulty score conveys exactly the same information as a number correct score in a simple fixed step size pyramid (Lord, 1970; Larkin & Weiss, 1974), since both are representative of the testee's path through the pyramid.

While the average difficulty score gives scores with more possible values than some scoring methods, e.g., difficulty of terminal item, it has a major limitation as a measure of individual differences in ability level. Since the average difficulty score includes a number of items near average ability for all testees, the score will not adequately reflect ability levels near the extremes of the ability distribution. For example, consider the five-stage pyramid in Figure 12a. The testee of highest ability would answer all items correctly and, therefore, would be routed directly from item 1 to item 15. The (n+1)th item score would put him in the upper 10% of the ability distribution, having correctly answered an item of .10 difficulty. Average difficulty score, however, would use the difficulty of all items answered, except the first (items 3, 6, 15 and "21") resulting in a score of .25.

Average difficulty score, as proposed by Lord, preserves the ordinal properties of other scoring methods while yielding more possible score values. However, it loses the direct interpretability of scores in terms of the testee's maximum level of performance that is derivable from the other methods. On the other hand, because the average difficulty score makes use of all the data in a testee's response record it might provide more stable ability estimates than those scoring methods based only on a single item.

The "average difficulty of all items answered correctly" score might eliminate the major deficiency of the average difficulty score. Since the "average difficulty correct" score eliminates from its computation those items answered incorrectly, it would be more interpretable as a direct measure of level of ability. This score would be more similar to the terminal item or maximum difficulty scoring methods except that it would be lowered somewhat for testees of high ability by the inclusion of the difficulties of the items used to route testees to their ability level. On the other hand, it would likely be more stable than the latter methods since it uses more information than the

terminal or highest difficulty scoring methods. However, the average difficulty correct score, since it would include items answered correctly, might be more susceptible to guessing than the other average difficulty methods, thereby lowering its reliability.

Hansen (1969, pp. 211-213) has proposed a method he calls "all item scoring." This method was proposed to permit the calculation of internal consistency reliability estimates of pyramidal tests for a group of testees. Essentially, Hansen proposes to score the pyramidal test "as if" each testee had responded to each item. Hansen assumes that item difficulty and ability are on the same continuum. The "all item" score assigns 2 points to each correct item and to items easier than that item at a given stage, 1 point to the next more difficult item at that stage, and 0 points to all items higher in difficulty. If an item at a given stage was answered incorrectly, that item and each more difficult item at that stage receives a 0. The item just below the incorrect item in difficulty increments the score by 1 point, and all other less difficult items result in 2 point increments. Hansen (1969) gives specific examples of this scoring method. Research by Larkin and Weiss (1974), however, shows that this scoring method gives results that are correlated .99 with the average difficulty score.

One last scoring method is a variation of the "final difficulty" methods described above. Rather than using the actual difficulty of the final item, the (n+1)th item, or other variations of this basic scoring procedure, these methods simply use the ordinal rank of these difficulties. Such ranked difficulties have been used by Bayroff and Seeley (1967) and Hansen (1969). These methods, of course, have the limitations of the methods they are based on, in terms of the limited number of score categories, and in addition lose the more direct interpretability of the parent scoring methods.

Most of the scoring methods proposed for the basic fixed branching, equal step size pyramids are applicable with little or no change to the more complex pyramidal models. Considerable research remains to be done on determining the differential utility of these scoring methods and on developing new scoring methods with different characteristics.

Advantages and limitations. In comparison to the two-stage models, pyramidal models appear to offer the advantage of measuring a wider range of abilities with considerably fewer items. While the two-stage test requires a routing test of, say, 10 items prior to administration of the measurement test, which increases the total number of items adiminstered, pyramidal tests use all items in the routing procedure and the measurement procedure simultaneously. Thus, pyramidal tests have the capability of estimating a testee's ability in as few as 10 or 15 items, approximating the number that a two-stage strategy

requires for a reliable routing test. Closely related to this advantage of the pyramidal model is the fact that the pyramidal models have the potential of covering a wider range of ability than the two-stage models, since two-stage models concentrate the measurement tests at a more restricted number of ability levels. To achieve equal coverage of the potential range of abilities, the two-stage strategy would require a considerably larger number of items than any of the variations of the pyramidal strategy. Thus, in general, the pyramidal strategies place less heavy demands on item pools than do the two-stage strategies. While a 10-stage pyramid requires 55 items, a two-stage test with a 10-item routing test and four 20-item measurement tests would require 90 items.

On the negative side, pyramidal testing models have two apparent logical disadvantages. First, pyramidal models have a "recoverability" problem not unlike that of two-stage models. Thus, pyramidal test scores are affected by chance successes, if guessing is a possibility, or by occasional incorrect test responses resulting from factors irrelevant to measured ability (e.g., errors in responding or inattention). Therefore, if a testee of very high ability answers one item incorrectly as a result of inattention, he is removed from the path leading to the highest test score and has no possibility of complete recovery to that highest score. On the other hand, the testee of very low ability who answers one item correctly by chance will not be routed to the lowest possible score but will instead receive an artificially higher score due to the one lucky guess. While the recoverability problem in pyramidal models is certainly not as serious as that of the two-stage model, this drawback may limit the utility of pyramidal tests in certain situations where scores of very high reliability are required.

Pyramidal models have one additional serious limitation. One objective of adaptive testing is to permit the tester to administer items that converge upon a difficulty level that is appropriate for each testee. To accomplish this goal, several writers (e.g., Hick, 1951; Lord, 1970, 1971a,c,d,e) have suggested that items should be administered at a level of difficulty where the testee answers about 50% of the items correctly, since it is these items that provide maximum "information" about the testee's ability. Pyramidal testing models may accomplish this goal for testees of about average ability; response records for these testees will show an approximate alternation between items answered correctly and those answered incorrectly. As the testee's ability deviates from the average difficulty of the pyramid, however, more test items are used to route the individual to his appropriate difficulty level, and fewer are available at the difficulty level which provides most information per test response. At the extremes of ability in a given pyramid, all of the items completed by a testee are used for routing, and none is available to indicate convergence on an appropriate difficulty level. In these cases the proportion of items answered correctly by the

testee approaches 0 or 1, in comparison to the more desirable .50 suggested by test theory and information theory.

Figure 5 illustrates this problem in the pyramidal strategy. The path followed by the solid line shows an apparent convergence on an ability level in the range of difficulty p=.40 to p=.35; in that range of the pyramid the testee approximately alternated between items answered correctly and incorrectly. The dotted line shows a downward path that appears to begin to converge around difficulties .75 to .80 (items 30, 38, 48), but there are simply not enough items available in the pyramid to determine if convergence has occurred. In comparison, however, the dashed line shows absolutely no evidence of convergence. Thus, the testee whose path through the test is traced by the dashed line continues to answer items correctly, having answered only one incorrectly, with no sign of convergence at the 10th stage. For this testee, the pyramidal strategy simply indicates that the testee is of high ability, but it cannot indicate how high the testee's ability really is. While a pyramid consisting of more stages would likely permit more testees to converge on an appropriate region of the item pool, larger pyramids make heavy demands on an item pool. Regardless of the number of stages in the pyramid, however, the determination of convergence for testees of high and low ability will always be based on smaller numbers of items, with the number of items decreasing with increasing distance from the center of the pyramid. The result will be ability estimates of lower precision for those testees.

Research issues. Pyramidal testing models are obviously a fertile field for both applied and theoretical research. There are a variety of questions to be answered with regard to the reliability (i.e., stability), validity, and utility of the testing models for estimating ability status of individuals. Thus, research is needed to compare the fixed step, decreasing step, truncated, multiple item, and differential response option branching pyramids on practical criteria. Within each of these models, however, there are a number of issues to be studied. These include study of the relative reliabilities and validities of the various scoring methods, the effect of varying step sizes and offsets on practical psychometric criteria, and the minimum number of stages required to construct a pyramid of maximum efficiency.

Other research questions include the effects of different ability distributions on the accuracy of ability estimates derived from the different pyramidal strategies and the psychological effects (e.g., in terms of such variables as test anxiety and motivation) of the different pyramidal structures, and their variations.

The Flexilevel Test

Lord (1971b) proposed the flexilevel test as a paper and

pencil technique for adapting test items to the ability level of the testee. The flexilevel test requires answer sheets that inform the testee of whether his response to each item is correct or incorrect. On the basis of this information, and in conjunction with the instructions presented at the beginning of the test, the testee branches to the next item in the test.

Although it is based on somewhat different logic than that of the pyramidal models, the flexilevel test can be viewed as a modified pyramidal adaptive test. Figure 13 illustrates the item structure for a flexilevel test. As Figure 13 shows, the flexilevel test consists of one item at each of a number of equally spaced difficulty levels. The flexilevel item structure is different from the typical pyramidal models in that the pyramidal models have more than one item at each difficulty level (with the exception of items at the extremes) while the flexilevel test has only one item at each difficulty level. Item 1 in Figure 13 is an item of approximately median difficulty (p=.50). The even-numbered items decrease in difficulty with increasing distance from the median difficulty item, while the odd-numbered items increase in difficulty.

Because the flexilevel test has only one item at each difficulty level, its branching rule differs from those of the pyramidal models. In the typical pyramidal test, an incorrect answer leads to a slightly easier item (e.g., an item one step lower in difficulty) while a correct response leads to a slightly more difficult item. This is possible in the pyramidal structure because there are a number of esentially equivalent items available at each level of difficulty. Which of these items is administered to a given testee depends on the testee's response pattern. In the flexilevel test, however, the branching rule states that following a correct response the next item given is the item next higher in difficulty which was not previously administered. And, following an incorrect response, the testee receives the item next lower in difficulty that has not been previously administered. The following examples will clarify the operation of this branching rule.

Figure 14a illustrates the path through a flexilevel test for a testee of relatively high ability. All testees begin with item 1, an item of median difficulty. A correct answer leads to an item of higher difficulty, just as in the pyramidal models. Thus, the testee depected in Figure 14a correctly answered items 1, 3, 5, 7, 9 and 11; as a result, he received items of higher and higher difficulty, moving from an item at p=.50 to an item at p=.20. At item 13 the flexilevel and pyramidal models diverge. Item 13 was answered incorrectly. Whereas in the pyramidal model the testee would next be administered an item of slightly lower difficulty (say p=.25), that item (item 11) had already been administered. The next less difficult item not already administered was item 2, with difficulty p=.55, since items of difficulty .50 to .25 had

Figure 13
Items Structure for a Ten-stage Flexilevel Test

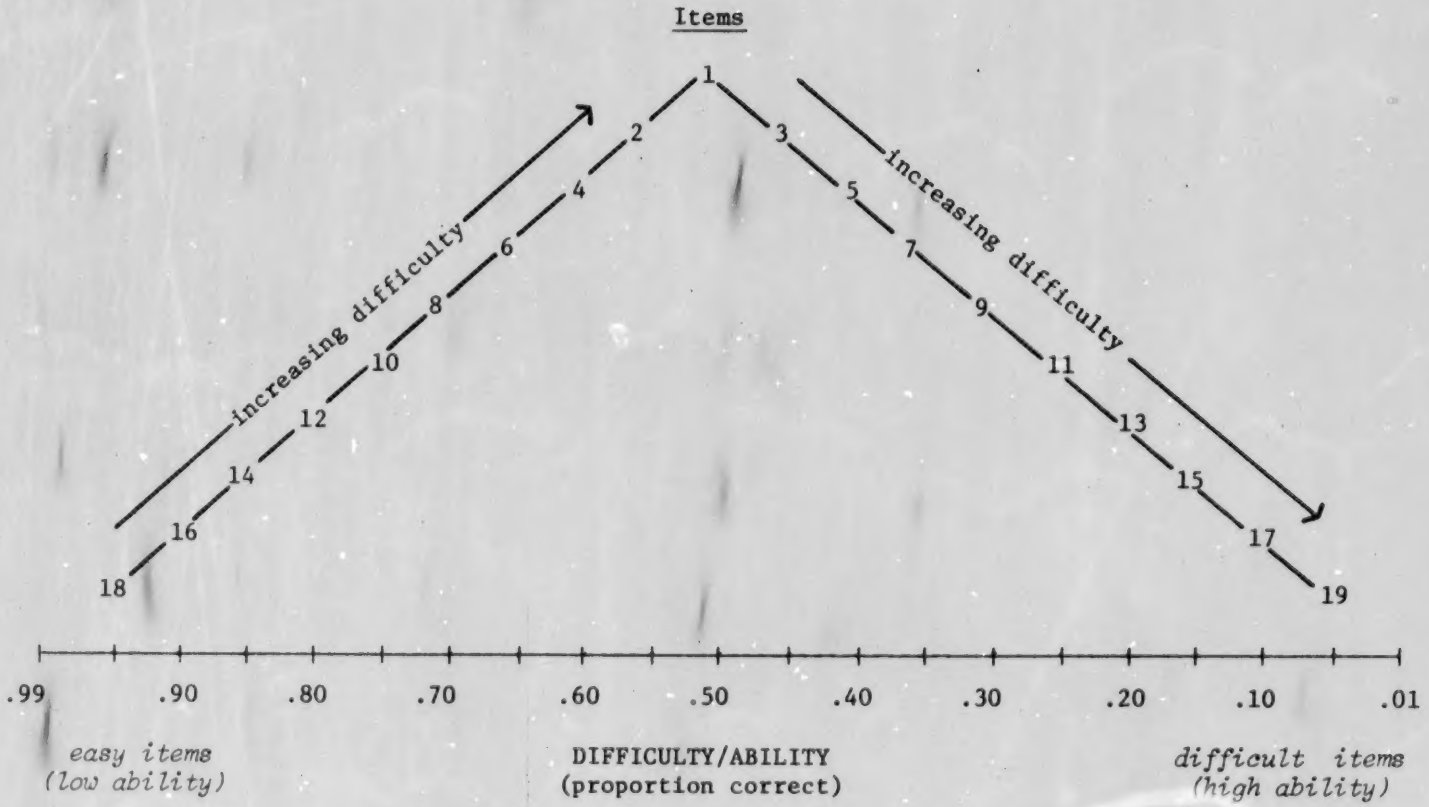
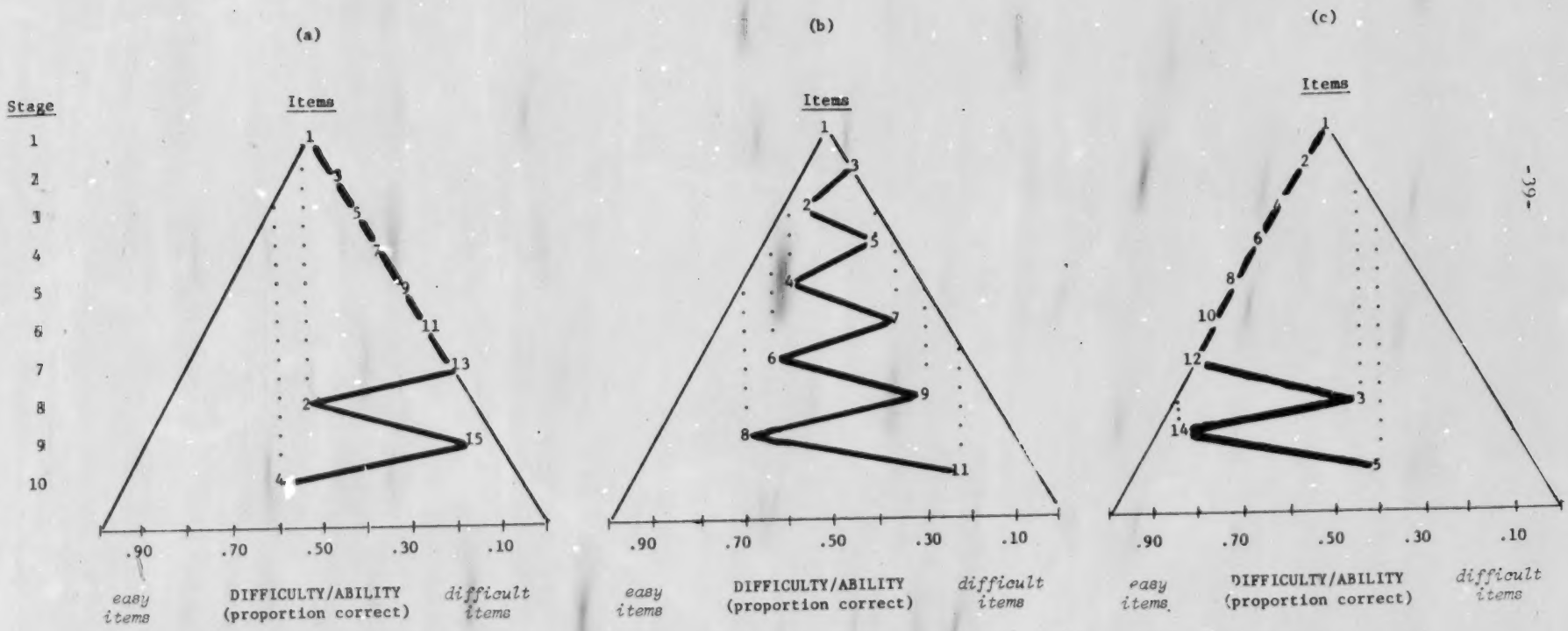


Figure 14
Sample Paths through a Ten-stage Flexilevel Test



already been administered. Therefore, under the flexilevel branching rule, item 2 was then administered and was answered correctly. It should be noted that item 2 is a quite easy item for a testee of high ability.

The flexilevel branching rule following a correct response is to administer the next more difficulty item not previously administered. Since the odd-numbered items 1 through 13 had already been administered, the next more difficult item not previously administered was item 15, of difficulty p=.15. That item was answered incorrectly, leading to item 4 (p=.60), the next less difficult item not previously administered. Again, item 4 is a very easy item for a high ability testee.

The flexilevel test terminates when the testee has answered half the available items (excluding the first item) in the pyramidal structure. Or, in other words, a 10-stage flexilevel test, in which each testee is to answer only 10 items, requires only 19 items in the item structure. In general, an n-stage flexilevel test (where n is the number of items to be answered by any testee) requires 2n-1 items distributed across the potential ability range with one item per difficulty level.

Following a flexilevel branching rule, a testee of average ability moves through the item structure alternating between successively more difficult and successively less difficult test items (see Figure 14b). Following a correct response to item 1 (p=.50), the testee in Figure 14b received item 3 (p=.45), which was answered incorrectly, leading to item 2 (p=.55). The alternation of correct and incorrect responses leads to items at p=.40, .60, .35, .65, .30, .70 and .25, in that order. Thus, the step size between items at successive stages gets larger and larger as the testee proceeds through the flexilevel test, alternating between items of increasing difficulty and those of decreasing difficulty. Since the testee depicted in Figure 14b is of average ability, the odd-numbered items (except item 1) are too difficult for him and the even numbered items too easy.

The response record of a low ability testee is diagrammed in Figure 14c. Items at stages 1 through 7 (items 1, 2, 4, 6, 8 and 10) were answered incorrectly by this testee, moving him from the item at p=.50 to item 12 at p=.80. The stage 7 item (item 12) was answered correctly, thus branching the testee to the next more difficult item previously unadministered, item 3 (p=.45), which was answered incorrectly. Item 14 (p=.85) was administered next and answered correctly, which finally branched the testee to item 5 at p=.40. At that point the testee had answered 10 items out of the 19 available, and testing was terminated. Again, items 3 and 5 are items which are inappropriate for this testee since they are considerably above his ability level.

The diagrams in Figure 14 illustrate how the flexilevel test adapts the range of item difficulties to individual differences in ability level. In essence, under the flexilevel strategy, each individual testee answers items from half the structured item network, in the general region of his ability level. The high ability testee (Figure 14a) answered the 50% of total items available in the range p=.15 to p=.60; items outside this range were not presented to that testee. The average ability testee answered the 50% of the total item structure in the range of p=.25 to p=.70, and the low ability testee answered a different 50% of the test items in the range of p=.40 to p=.85. The flexilevel procedure does identify a region of the item pool of approximately appropriate difficulty for each individual. But, after that appropriate difficulty level is reached, the remaining items administered provide little information on the testee's ability level since they are either too difficult or too easy for the testee.

Scoring. The flexilevel test is scored by counting the number of questions answered correctly. Lord (1971b) shows that the flexilevel strategy is designed so that all testees who obtain a given total correct score have answered the same subset of items. In other words, testees with the same total score have been presented with the same items and have answered the same total number of those items correctly, but have not necessarily answered the same items correctly or incorrectly. Thus, the pattern of responses (in terms of which items were correct and which were incorrect) can vary even though the total score of two testees is the same. In an effort to further distinguish among the different paths leading to a given score, Lord proposes that an additional half point be added to the scores of each testee ending the flexilevel test with an incorrect answer; testees who answer the last item correctly do not receive the half-point bonus. The rationale for this scoring method is discussed by Lord (1971b, p. 150-151).

Advantages and limitations. The flexilevel test has a number of advantages over competing adaptive testing models. Like the two-stage test, it might be possible to administer a flexilevel test by paper and pencil. The flexilevel test is easy to score, and scores are relatively easy to interpret since all testees who obtain the same score have answered the same items, i.e., "taken the same test." However, as in a conventional test, the same total score does not mean that two testees answered the same items correctly and incorrectly. A final major advantage is its item economy; of all the adaptive strategies proposed, the flexilevel test requires the smallest item pool. While a 10-stage pyramidal test requires a 55-item structure, a 10-stage flexilevel test requires only 19 items.

The flexilevel strategy has several potential limitations. The number of items administered to each testee is the same. The result is that the item difficulties diverge from the testee's

ability level, as described above. Such divergence might have detrimental effects on testee motivation and result in guessing behavior on items that are too difficult. Secondly, the flexilevel test includes only one item at each level of difficulty. The result of this structure might be not enough items at any given difficulty level to accurately determine a testee's ability status with a high degree of precision. Thus, in practical terms, scores derived from flexilevel testing might be more unstable than those derived from other adaptive strategies, particularly if guessing is possible.

Research issues. Since the flexilevel test was proposed as a paper and pencil procedure, a relevant question is whether it can be practically implemented in that mode or whether automated administration is necessary. Research should also be conducted on methods of scoring the flexilevel test. Although Lord makes a convincing case for the "half-point bonus," the possibility of other, more meaningful, scoring methods still exists since some scoring methods might have more utility than others in certain applied situations. Another research question concerns the maximum and minimum effective lengths of the flexilevel testing procedure. How many items should be administered to each testee to yield scores with certain desirable characteristics? Is a 60-stage flexilevel test, as studied theoretically by Lord (1971d), really necessary? Is a 15-stage flexilevel (Stocking, 1969) too short? Research also needs to be conducted on the effectiveness of different step sizes in the flexilevel test. Lord studied only a 60-stage flexilevel test, varying step size but holding item discriminations constant. An adequate investigation of the flexilevel procedure would require that these parameters be varied in empirical studies to supplement the suggestions derivable from theoretical analysis.

Lord (1971b, p. 150) claims that "exact determination of difficulty levels is not necessary for proper comparison among examinees." Logically, however, it appears that inexact difficulties will result in only a rough ordinal scaling of the testees, rather than the equal interval scaling more generally desirable. Thus, empirical research is needed on the effects of variations of item difficulty estimates on the utility of flexilevel scores. Although Lord has, under specific theoretical assumptions, studied the effects of random guessing on the flexilevel test, no empirical data from live testing is available on the effects of guessing. It could be hypothesized that, because the flexilevel test approximately adjusts the difficulty of test items to the ability of the testee, guessing should be reduced in comparison to conventional tests. However, since item difficulties in the flexilevel test diverge from the testee's ability level near the end of the test, guessing might increase, in comparison to other adaptive strategies where item difficulties converge on the testee's ability level.

Finally, the psychological impact of the flexilevel test

needs to be studied. Lord (1971d) suggested that the effect of item tailoring on the testee's attitude and performance be studied. While this question is appropriate to all adaptive testing strategies, one further question is relevant to the flexilevel test. As has been indicated above, the flexilevel test, in contrast to all other adaptive testing strategies, is a non-convergence procedure. In the flexilevel procedure, once the region of difficulty appropriate to the testee is approximately located, the step size increases rather than decreases. The net effect is that as the flexilevel test proceeds through successive stages, the testee is administered a series of items which tend to alternate between items that are much too easy for him and items that are much too difficult. These items provide very little or no information (Hick, 1951; Lord, 1971a) on a testee's ability level, since the probability of getting them correct is close to 1.0 or 0.0. And, just as in the other pyramidal models, the overall appropriateness of item difficulties for an individual is inversely related to the distance of his ability level from the median ability level.

This divergence procedure might have psychological effects on the testee. Assuming that the testee has some subjective feeling of whether he answers an item correctly or incorrectly, the flexilevel strategy can be viewed from the standpoint of reinforcement theory. Once the divergence procedure has begun, following a correct response the testee is administered a difficult item well above his ability level. Since the item is too difficult, he is likely to answer incorrectly. His "reward" is an easier item, which he is likely to answer correctly. For answering an item correctly, however, the testee is "punished" by receiving another very difficult item. Thus, each correct response is "punished" and each incorrect response is "rewarded," with the "intensity" of the "reward" or "punishment" increasing as the testee progresses through the flexilevel test. Once the testee realizes what is happening, it might be natural simply to respond with incorrect answers, in order to obtain easier items which he knows he can answer correctly. The testee might not answer these easier items correctly, however, because of the perceived "punishment" of a quite difficult item which might follow a correct response.

While this phenomenon might occur in the pyramidal models, since the direction of branching is the same, because the pyramidal branching is between contiguous difficulty levels the change in item difficulty might be less likely to be perceived by the testee. In any event, the potential effect is worth investigating, as is the probable effect on guessing behavior resulting from the administration of items of increasing difficulty, well above the testee's ability level, that is characteristic of the flexilevel strategy.

The structure of the flexilevel test might also serve to induce frustration in low ability testees, in an analagous fashion

as might conventional tests and the other pyramidal models. Since each testee begins the test at items of median difficulty, the very low ability testee must answer incorrectly a number of items which are too difficult for him before he reaches items which he can answer correctly. Assuming, again, that the testee maintains a subjective account of his performance, he may become frustrated before he reaches items that are easy enough for him.

The Stradaptive Test

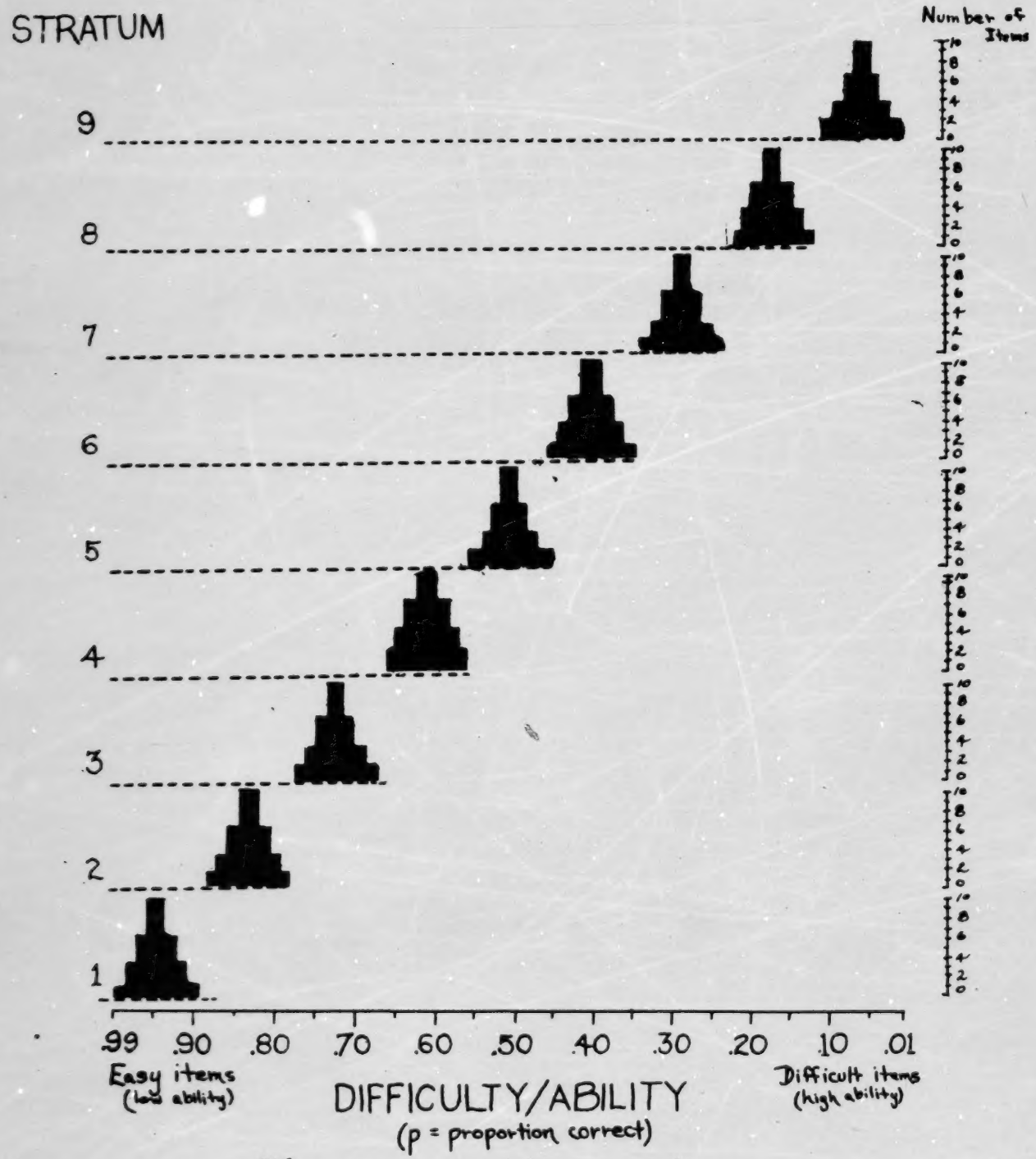
The stradaptive (stratified-adaptive) computerized test (Weiss, 1973) operates from an item pool in which test items are grouped into levels, or strata, according to their difficulties. Thus, the stradaptive item pool might include nine strata of items. Each stratum can be thought of as a peaked test in which the items are clustered around some average difficulty level. The strata are arranged in increasing order of difficulty.

Figure 15 gives a diagrammatic representation of the structure of a stradaptive item pool. In Figure 15, the height of the distribution within each stratum represents the number of test items available in a given difficulty interval. Each successive stratum includes a subset of items within a narrow range of difficulty. The mean difficulty of each stratum is higher than the preceding stratum. Stratum 1 is the easiest of nine peaked tests, consisting of items in the range of difficulty from .89 to .99. Most of the items in stratum 1 have difficulties around .94, the mean difficulty of the stratum, or the point at which that subset of items is peaked, while a few of the items have difficulties as high as .99 and as low as .89. Stratum 2 is a peaked test slightly more difficult than that of stratum 1. The average difficulty of stratum 2 is about .83; most of the item difficulties in that stratum are clustered around .83, and a few have difficulties as extreme as .78 and .88. The most difficult stratum (number 9) consists of items ranging from .11 to .01 in difficulty, with the average difficulty of the items at about p=.06.

Unlike the other fixed branching models, stradaptive testing begins with the estimation of an individual's ability level from either prior information available on the testee or from his self-report (e.g., Weiss, 1973, p. 16). This information determines the testee's "entry point" into the hierarchy of strata of increasing difficulty. Based on whatever prior information is available, the testee of lower estimated ability begins the stradaptive test with easier items, and the testee of higher estimated ability begins with more difficult items.

Branching in the stradaptive test occurs between strata, and any of the branching rules (e.g., up-one/down-one, up-one/down-two) can be used. In the case of the up-one/down-one rule,

Figure 15
Distribution of Items, by Difficulty Level, in a Stradaptive Test



a correct answer to an item at one stratum leads to the next available item at the stratum next highest in difficulty. An incorrect answer to an item at a given stratum branches the testee to the next available item at the stratum next lower in average difficulty. Thus, branching from item to item is not fixed by the branching rule, since no item from the same set of items invariably follows a response to any given test item. Rather, branching in the stradaptive test is from stratum to stratum. The item to be administered at each stratum is the first of the remaining items not previously administered at that stratum. In stradaptive testing, the step size and offset refer to the average difficulties of the strata and not of single items.

The stradaptive test can be viewed as a search for the peaked ability tests, out of those available in the stradaptive item pool, which provide meaningful information on a testee's ability level. Items which are too easy or too difficult, provide little or no information on a testee's ability level. Thus, the stradaptive branching procedure is designed to converge upon the region of the item pool of appropriate difficulty for a given testee. In the process of this convergence, the stradaptive test will locate a stratum of the item pool at which the individual answers all (or almost all) of the items correctly; this can be referred to as the "basal stratum." At the same time the procedure will locate a "ceiling stratum," or a stratum at which the individual answers all (or almost all) the items incorrectly. In between those two strata the testee will answer about 50% of the items correctly (assuming minimal guessing).

A unique feature of the stradaptive strategy as compared to the other fixed branching strategies (although characteristic of the variable branching strategies) is that the number of items administered to a given testee is not determined in advance. Rather, the number of items to be administered to each testee is free to vary and depends on how quickly a given termination rule or criterion is reached. Given that a sufficient number of items has been administered to identify a reliable "ceiling stratum," the stradaptive test can be terminated. Where guessing is a possibility, the definition of a ceiling stratum (and, therefore, the termination rule) can explicitly take guessing into account. With 5-alternative multiple choice test items, random guessing will result in correct answers to about 20% of the items at any given stratum. Therefore, for this type of test item, the ceiling stratum can be defined as that difficulty stratum at which an individual answers 20% or less of the items correctly, assuming that the decision is based on, say, five items. Once the ceiling stratum has been located for a given testee, the stradaptive testing procedure can be terminated. Further details on the rationale and construction of the stradaptive tests are in Weiss (1973).

Figure 16 shows a typical response record from stradaptive testing. Based on his own estimate of his ability level, the testee began the stradaptive test at stratum 5. Stratum 5 includes those items in the stradaptive pool of average difficulty. His response to item 1 was correct (1+) thus branching him to the first available item at stratum 6, a peaked test composed of items slightly more difficult than those at stratum 5. The second item administered was answered correctly (2+), branching the testee to the first available item at the stratum next highest in difficulty, stratum 7. Another correct answer led the testee to stratum 8 for the first item of that stratum, which he answered incorrectly (4-). A series of alternating correct and incorrect responses through the eighth item kept him between strata 7 and 8. An incorrect response at stage 9 returned him to stratum 6 and thereafter he alternated principally between strata 6 and 7. Finally, the 19th item administered (which was the ninth item available at stratum 7) was answered correctly, and the testee was branched again to an item at stratum 8, which he answered incorrectly (20-). At that point it was determined that he had reached his ceiling stratum; he had answered five items at stratum 8 and none was answered correctly. Testing was then terminated.

As Figure 16 shows, the stradaptive testing procedure identified a ceiling stratum (stratum 8, with a proportion correct of 0.0) and a basal stratum (stratum 6, with a proportion correct of 1.00). In between these two strata was a peaked test whose items gave meaningful information on the testee's ability (stratum 7, at which he answered 56% of the items correctly). That the entire testing procedure was appropriately adapted to the testee's ability level is further supported by the total proportion correct of .55, reflecting the fact that on the average the 20 items administered gave almost optimal (p=.50) average information on his ability level.

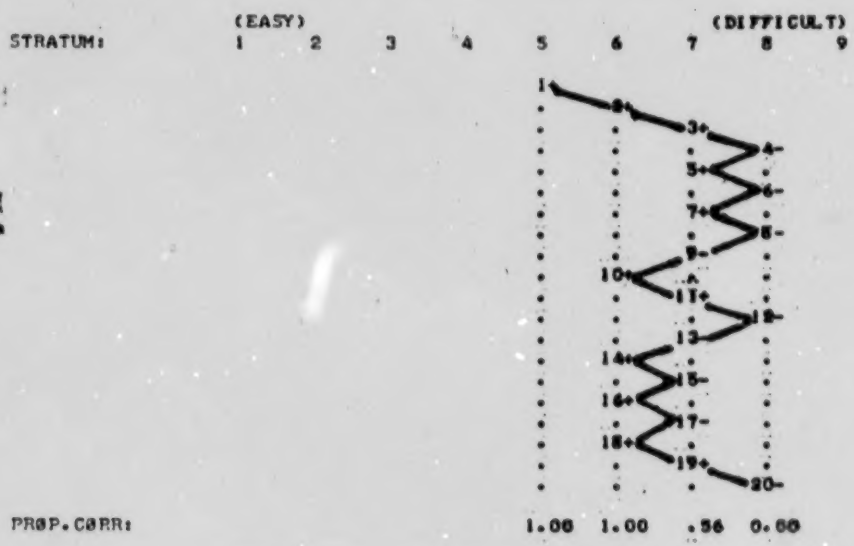
A second example of stradaptive testing is shown in Figure 17. Based on previous information about the testee's probable ability level, he began the stradaptive test at stratum 8. His first answer was correct, leading him to stratum 9 for the next item. Following item 2 he answered all but one of the next six items incorrectly. The ninth item administered was the first item available at stratum 4, which was answered correctly. The response record then shows a series of large fluctuations between strata 5 and 8, with one additional item answered correctly at stratum 4 and one answered incorrectly at stratum 9. At item 36, the responses finally began to alternate between correct and incorrect to items at strata 7 and 8. Finally, at item 41 ten items had been answered at stratum 8 with only two answered correctly. Having thus met the termination criterion of 20% or fewer correct, the stradaptive test was terminated after 41 items.

The test record in Figure 17 again illustrates the basic

Figure 16
Report on a Stradaptive Test for a Consistent Testee

REPORT ON STRADAPTIVE TEST

ID NUMBER: DATE TESTED: 73/07/12



TOTAL PROPORTION CORRECT= .550

SCORES ON STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 1.49
2. DIFFICULTY OF THE N+1 TH ITEM= 1.44
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 1.49
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= 1.33
5. DIFFICULTY OF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 1.33
7. INTERPOLATED STRATUM DIFFICULTY= 1.37
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= .80
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= 1.28
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= 1.28

Consistency Scores

11. SD OF ITEM DIFFICULTIES ENCOUNTERED= .59
12. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY= .46
13. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY BETWEEN CEILING AND BASAL STRATA= .18
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 1.36
15. NUMBER OF STRATA BETWEEN CEILING AND BASAL STRATA= 1

Figure 17
Report on a Stradaptive Test for an Inconsistent Testee

REPORT ON STRADAPTIVE TEST

ID NUMBER: [redacted] DATE TESTED: 73/07/02

TOTAL PROPORTION CORRECT= .488

SCORES ON STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 1.89
2. DIFFICULTY OF THE N+1 TH ITEM= 1.01
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 1.53
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= 2.01
5. DIFFICULTY OF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 1.33
7. INTERPOLATED STRATUM DIFFICULTY= 1.36
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= .72
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= .76
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= 1.24

Consistency Scores

11. SD OF ITEM DIFFICULTIES ENCOUNTERED= .86
12. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY= .74
13. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY BETWEEN CEILING AND BASAL STRATA= .50
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 2.64
15. NUMBER OF STRATA BETWEEN CEILING AND BASAL STRATA= 3

characteristics of stradaptive testing. For that testee, stratum 4 was identified as the basal stratum or the stratum at which the testee answered all items correctly. Stratum 8 was the ceiling stratum, the stratum at which he answered at or below chance expectation. In between these two strata, the proportions correct were between 1.00 and 0.00, with the total proportion correct for the testee (p=.488) near .50. Thus, for this testee, as for the testee shown in Figure 16, the stradaptive testing strategy located the region of the item pool in which testing provides most average information per item for a given individual.

It is interesting to draw some contrasts between Figure 16 and Figure 17. First, the number of items administered to the testees differed; the first required only 20 items while the second required 41 items to reach the termination criterion. This illustrates a characteristic feature of the stradaptive test. That is, that the number of items to be administered to any testee is not determined in advance, but is a joint function of the appropriateness of the entry point, the testee's unique response record, and the termination criterion. Secondly, the response records of Figures 16 and 17 differ in the number of strata used in testing. In Figure 16, the one item administered at stratum 5 served only to route the testee to strata 6 through 8, where all testing was concentrated. Thus, in Figure 16 testing was effectively carried out at only three strata. In contrast, the response record in Figure 17 utilized items in strata 4 through 9, with testing concentrated in 4 strata, strata 5 through 8. The responses of the first testee showed a narrow oscillation between strata 6 and 8; on the other hand, the second testee's response record seemed to fluctuate in a wide oscillation between strata 5 and 8. The first testee was thus more consistent in his responses than was the second. For the second testee it appears that three of the peaked tests comprising the strata were appropriate: stratum 5, p=.60; stratum 6, p=.67; and stratum 7, p=.54. In contrast, for the first testee only stratum 7 (p=.56) provided information on his ability level.

Scoring. There are a number of possible ways of scoring the stradaptive test (Weiss, 1973). Some methods will likely have greater reliability, validity or utility than others; thus, the choice of the most appropriate methods will have to await the results of future research. Some of the scoring methods are borrowed from the logic of the pyramidal models, while others are unique to stradaptive testing. Figures 16 and 17 show scores for the two sample stradaptive test response records.

In the stradaptive test, ability level can be scored as:

1. Difficulty of the most difficult item answered correctly.

2. Difficulty of the (n+1)th item or the item that the testee would have answered next if testing had continued. This scoring method is appropriate since testing usually terminates with an item at the ceiling stratum. Thus, the testee whose last item was answered correctly would receive a higher score than the testee who answered the same last item incorrectly.

3. Difficulty of the most difficult item answered correctly at the highest non-chance stratum. For this scoring method the highest non-chance stratum is that stratum immediately below the testee's ceiling stratum. The ceiling stratum is the stratum at which termination occurred, and is that stratum at which the testee answers 20% or less of the items correctly, having completed five or more items at that stratum.

4. Difficulty of the highest, or most difficult, stratum at which an item was answered correctly. This scoring method, and methods 5 and 6, use the average difficulty of all items at a given stratum as the difficulty level, or score, for that stratum.

5. Difficulty of the stratum of the (n+1)th item. In contrast to method 2, this method uses as the score the average difficulty of the stratum at which the (n+1)th item would occur.

6. Difficulty of the highest non-chance stratum, i.e., the difficulty level of the stratum just below the ceiling stratum.

7. Interpolated stratum difficulty. This method uses the proportion correct at the highest non-chance stratum to interpolate the distance in difficulties to the next higher or lower stratum. The formula is:

$$A = \bar{D}_{c-1} + S(P_{c-1} - .50)$$

where A is the testee's ability score,

$\bar{D}_{c-1}$ is the average difficulty of the items at the $(c-1)^{th}$ stratum, where c is the ceiling stratum,

P_{c-1} is the testee's proportion correct at the $(c-1)^{th}$ stratum,

$S = \bar{D}_c - \bar{D}_{c-1}$ if $P_{c-1} > .50$

or $\bar{D}_{c-1} - \bar{D}_{c-2}$ if $P_{c-1} < .50$

and $\bar{D}$ is the average difficulty of the designated stratum.

Thus, this scoring method will give higher ability estimates for the testee who gets .80 of the items correct at the $(c-1)^{th}$ stratum in contrast to the testee who answers only .20 of the same items correctly.

8. Mean difficulty of all items answered correctly.

9. Mean difficulty of items answered correctly between (but not including) the ceiling stratum and the basal stratum. The basal stratum is that stratum at which the testee gets all (1.00) of the items correct.

10. Mean difficulty of all items answered correctly at the highest non-chance $[(c-1)^{th}]$ stratum.

A unique feature of the stradaptive test is its "consistency scores." These scores reflect the consistency of the interaction between the testee and the items. A consistent testee is one whose item response record shows small variability in the difficulties of the items encountered. An inconsistent testee is one who utilizes a large number of strata and for whom the convergence on strata of appropriate difficulty is less precise. The response record shown in Figure 16 is a consistent one while the response record in Figure 17 reflects an inconsistent response record. It is possible that appropriate measures of consistency could be used to develop an individualized "standard error of measurement."

There are a number of ways of quantifying consistency in the stradaptive test response record. Figures 16 and 17 show results for the following consistency scores:

11. Standard deviation of item difficulties encountered. This score will be lower for testees who use a smaller number of strata (e.g., Figure 16) and higher for those whose responses vary across more strata (e.g., Figure 17).

12. Standard deviation of the difficulties of the items answered correctly. Again, this value will be lower for the more consistent testees and higher for those less consistent. However, because items answered incorrectly are not considered in this score, and because incorrect answers will occur at the ceiling stratum, this score will be lower, in general, than the previous scores.

13. Standard deviation of item difficulties for items answered correctly between the ceiling and basal strata. This score attempts to correct for inappropriate

entry points into the stradaptive structure, which will artificially inflate the two previous scores. An entry point is inappropriate if it results in the administration of items below the basal stratum or above the ceiling stratum.

14. The difference in average difficulties between the ceiling and basal strata. Consistency is a function of the distance between the two strata on the difficulty continuum.

15. Number of strata between ceiling and basal strata. This score indicates the number of peaked tests (strata) which are necessary to provide information on each testee's ability level.

These represent scoring methods for the stradaptive test that have been proposed thus far; further discussion of their rationale and characteristics can be found in Weiss (1973). Other methods of scoring the stradaptive test are likely to be developed in the future as experience is gained with the results of stradaptive testing of different populations.

Advantages and limitations. Stradaptive testing has several advantages over the two-stage and the other multi-stage fixed branching models. First, the termination rule currently under investigation is explicitly designed to take account of guessing, although it should work equally as well when guessing does not occur, as in a test using free-response items. Second, the stradaptive test can be completely recoverable. That is, if a testee answers earlier items correctly or incorrectly by chance, he will still be able to obtain maximally high or low scores on the test, by some scoring methods. Such complete recovery is not possible in the pyramidal and flexilevel strategies, and the two-stage strategy frequently results in routing errors. A third advantage of the stradaptive test is that rather than assuming that a given item pool is appropriate for a testee, as does the pyramidal strategy, the stradaptive test locates the region of the item pool which provides most information on a testee. Thus, most testees will obtain a total proportion correct of about .50 on the stradaptive test. Fourth, in stradaptive testing all individuals do not begin the test with the same item. Thus, testing makes use of prior information which should act to reduce the number of items to be administered to each testee. A fifth advantage of the stradaptive test is that it permits the number of items administered to each testee to vary. As a result, the precision of stradaptive test scores can be held constant, to some extent, across individuals by continuing testing until the required degree of precision is reached. Since the stradaptive item pool includes adequate numbers of items for all ability levels, it should permit measurements with relatively equal precision at all ability levels, including the extremes. A sixth advantage

of the stradaptive test is the possibility of computing consistency scores which might function as individual "errors of measurement." Such consistency scores should be related to the stability of measurement for given individuals, thus permitting more accurate longitudinal predictions for those testees. Finally, although the stradaptive test was designed for computer administration, it would be possible to administer it using a testing machine especially designed for that purpose.

Research issues. Since little is known about the characteristics of the stradaptive testing strategy, a variety of research questions can be raised to identify its optimal properties. Stradaptive tests can use almost any of the branching rules that have been proposed for the pyramidal models, with the exception of the Robbins-Monro shrinking step size procedures. Thus, a relevant research issue is the optimal step size in terms of the distance between contiguous strata. In addition, studies need to be done on the optimal number of strata, and the optimal (minimum and maximum) number of items at each stratum. Research should also be conducted on the "offset," or the number of strata spanned by each branching decision, since stradaptive tests can also use variable offset rules such as "one-up/down-two."

The variety of methods of scoring the stradaptive test also leads to research questions such as which scoring methods give the most stable results, which most validly reflect the testee's ability level, and which best predict external criteria.

The termination rule is a critical aspect of the stradaptive strategy. Although one logical termination rule has been proposed, a variety of others are possible. For example, testing might be terminated after a specified number of items have been administered or when the total proportion correct tends to stabilize near .50. The ceiling stratum termination rule might be changed from .20 to another value more representative of true guessing behavior. Research should also be conducted on the psychometric and practical utility of individual testee data concerning the number of strata used in testing or the variance of item difficulties answered correctly, or encountered by, a given testee. More complex analyses of the stradaptive test record might employ maximum likelihood estimates of ability. at each stage of testing coupled with a termination rule that ends testing when the error of the ability estimate converges on a pre-specified value.

Variable Branching Models

The fixed branching models all have in common the fact that the branching rule (step size and offset) is determined in advance and applied to all responses of all testees. Thus, the item pool is structured prior to testing so that certain paths through the item pool will occur for specified sequences of

test responses.

The variable branching multi-stage models, on the other hand, do not operate from a structured item pool in which a correct or incorrect response to a given item will always result in the administration of specific items at the next stage. In the variable branching models, step size and offset do not exist. Rather, these models operate from an item pool calibrated by difficulty and discrimination. The general item selection rule is to choose that item, out of all remaining unadministered items, which best fits the requirements of the mathematical model being used. The two general kinds of models proposed to date include Bayesian and maximum likelihood approaches to adaptive testing.

Bayesian Strategies

The Bayesian strategies of adaptive testing are based on application of Bayes' theorem (e.g., Lindley, 1965; Schmitt, 1969; Winkler, 1972) to the sequential response processes of adaptive testing. In its simple discrete case Bayes' theorem allows calculation of the probability that an event E_1 has occurred, given that another event E_2 has occurred. This probability is known as the "posterior" probability of E_1. To calculate the posterior probability of E_1 using Bayes' theorem, one needs to specify the probability of E_1 ocurring (whether or not E_2 occurs), known as the "prior" probability of E_1, the probability of E_2 occurring given that E_1 has occurred, and the probability of E_2 occurring given that E_1 has not occurred. The last two probabilities are known as the "likelihoods" of E_2. In the continuous case of Bayes' theorem, which is used in adaptive testing models, probabilities become probability distributions, and likelihoods become likelihood functions.

The general procedure used in the Bayesian approaches to adaptive testing involves the following steps. First, a prior estimate of the testee's ability and the standard error of that estimate are made at each stage of testing, based on whatever information is available about the testee. Second, a test item is selected from the item pool which has been previously calibrated in terms of the difficulties and discriminations of the items. All items in the item pool which have not already been used in testing a given testee are considered as possible next items to be administered. This process identifies the one item in the pool that will most reduce the uncertainty of a testee's ability estimate if it is administered. The selected item is usually the item of difficulty closest to the testee's estimated ability level. Following administration of the selected item, the prior ability estimate and the information obtained from administering that item (i.e., it was answered correctly or incorrectly) are combined by means of Bayes' Theorem to obtain a posterior ability estimate. This latter estimate is a revised estimate based on what was known about

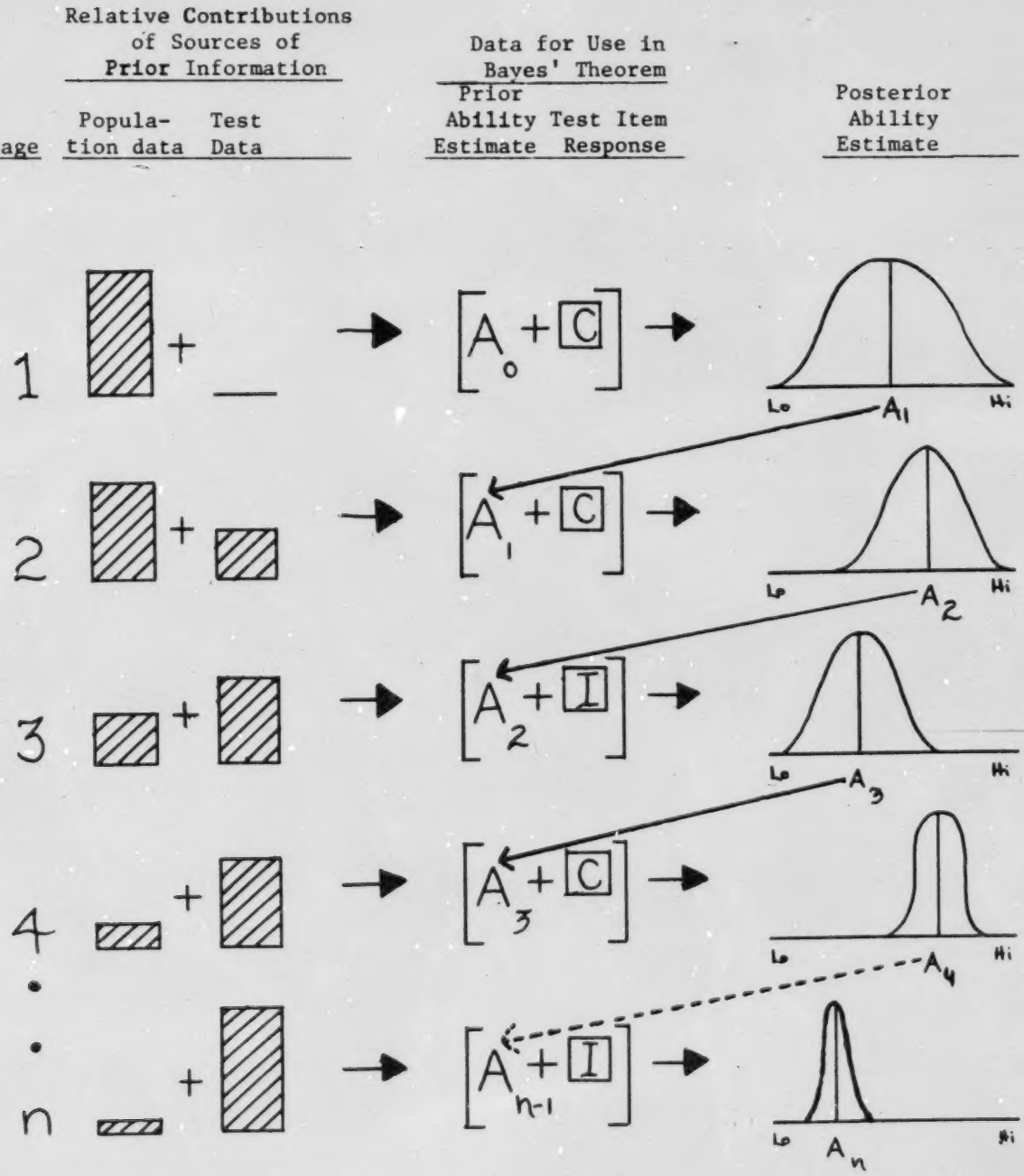
the testee's ability before the test item was administered and what was learned about the testee from administering the chosen test item.

Associated with the posterior ability estimate are data on its certainty, or the standard error of that ability estimate, a value which bears some similarity to a standard error of measurement. The Bayesian estimation procedure guarantees that the standard error of estimate will be reduced following the administration of each test item, since every item administered to a testee provides some information about his ability level, regardless of whether he answers it correctly or incorrectly. Based on the value of the standard error of the ability estimate, the tester can decide at each stage of testing whether to terminate testing or whether to continue. If the error of estimate has reached a sufficiently small value, testing can be terminated. If, however, the last posterior ability estimate has a larger error of estimate than the tester is willing to tolerate, the tester will likely decide to continue testing by administering another test item. If testing is continued, the posterior ability estimate from the previous stage of testing becomes the prior ability estimate for the next stage. A new test item is chosen according to the uncertainty minimization criterion, it is administered, and the Bayesian procedure uses that information in conjunction with the new prior ability estimate to obtain a new posterior ability estimate. The error of estimate of the new posterior ability estimate is computed, and the decision is made as to whether to continue or to terminate testing. This cyclical process is continued until a pre-designated degree of accuracy is reached. It is obvious that the number of test items to be administered to each testee is free to vary under the Bayesian strategies. If the number of items is determined in advance, the resulting standard errors of the ability estimates will vary among testees.

Two strategies have been proposed for the application of Bayes' Theorem to adaptive testing. The two strategies differ in the kind and extent of prior information used to estimate ability at each stage of testing, as well as in specific computational details.

Novick's strategy. Novick's (1969) approach, although it is not completely articulated and was proposed more as a theoretical model than as an operational testing strategy, has considerable logical appeal. Figure 18 is a schematic representation of Novick's proposal. Novick proposes the use of both population data and test item response data in determining the prior ability estimate for a testee. The heights of the shaded bars in Figure 18 indicate the relative contribution of these two types of data in determining the first prior ability estimate. As Figure 18 shows, at stage 1 of testing (i.e., before the first test item has been administered) test information provides no prior information on a testee's ability level. However,

Figure 18
Diagrammatic Representation of Novick's Bayesian Testing Strategy



data describing the population to which the testee belongs can be used as prior information. For example, in the ability domain being measured by an adaptive test, male testees will have a specified mean and variance of ability; female testees might have a different (higher or lower) mean and a different variance. Given the lack of other data on the individual (i.e., he/she has responded as yet to no test items), the mean and variance of ability in the population to which he or she belongs provides the best initial estimate of his or her ability level. Population data, then, is used to determine the prior ability estimate at the first stage of testing.

After the initial prior ability estimate has been determined, a test item is selected for administration to the testee. In Novick's approach, item order can be fixed or item selection can occur according to the optimization rules described earlier. The item is answered by the testee (C=correct; I=incorrect, in Figure 18), and a posterior ability estimate and its variance are computed, using Bayes' Theorem, from the prior information and the information obtained from the test response. The posterior ability estimate from stage 1 is then used as the prior ability estimate in stage 2. As the heights of the shaded bars of Figure 18 show, at the second stage of testing, population data still provide most of the prior information, but test data (based on the stage 1 item administered) begin to provide some prior information. At successive stages, 3 through n, the relative contributions of population data and test data reverse until, at stage n, population data provide little or no prior information, while test data provide the basis for almost all the prior information used in the computations derived from Bayes' Theorem.

Figure 18 also illustrates two other characteristics of the Bayesian testing procedure. First, following evaluation of each item response as correct or incorrect, a posterior estimate of ability is obtained. A correct response will lead to a higher ability estimate, while an incorrect response results in a lower ability estimate. Second, the procedures guarantee that the error of the ability estimate will be reduced at each stage of the testing procedure. Thus, the normal distributions shown in Figure 18, which reflect the standard error of the posterior ability estimates, become narrower at each successive stage of testing. At the nth stage of testing, the error of the ability estimate is quite small, indicating a very narrow range within which the true ability is probably located.

Owen's strategy. Owen's (1969) Bayesian testing model differs from Novick's primarily in that it does not use population data as prior information throughout the testing procedure. Owen uses non-test data only to obtain the first prior ability estimate. Thus, beginning with the second stage of testing, the prior ability estimates in Owen's model are based entirely on the test data provided by the testee. The stage 1 prior

ability estimate is based on whatever is known about the ability level of the testee. When there is no basis on which a differential ability estimate can be made, the stage 1 prior ability estimate can simply be set to arbitrary values, such as a mean of 0.0 and a standard deviation of 1.0. This would indicate that no prior information is available about the testee.

In a fashion similar to Novick's approach, the posterior ability estimate from stage 1 in Owen's model becomes the prior of stage 2. The stage 2 item is chosen from the entire item pool to minimize the variance of the stage 2 posterior ability estimate. Next, the stage 2 item is administered and the posterior ability estimate is calculated. The stage 2 posterior estimate itself then becomes the stage 3 prior ability estimate.

Owen's model also differs from Novick's in that it is completely articulated and operational, and includes a means of accounting for guessing as a function of the difference between testee ability and the difficulty of a given test item. Further, Owen's model selects items somewhat differently than Novick's.

Figure 19 shows the response record of an individual actually administered a Bayesian adaptive test, constructed according to Owen's model and using the item pool developed by McBride and Weiss (1974). Shown are a record of the ability estimate and its standard deviation at each stage of testing. At stage 0, the posterior ability estimate is based solely on the entry point (E) information. Stage 0 prior ability estimates were based on the testee's subjective evaluation of his ability level (see Weiss, 1973, pp. 15-16). Using that information, the testee obtained an ability estimate of z=-.85. The standard deviation of that ability estimate was z=1.87. The stage 1 item, selected from the item pool according to Owen's (1969) equations, was administered and answered correctly; the ability estimate increased to z=+.19 and its error reduced to 1.34. The stage 2 item was chosen, administered and answered incorrectly (0); and the ability estimate was lowered (to z=-.85) and its error decreased to .94. The sequential procedure continued with the ability estimate slowly converging on a value of -1.35 standard deviations below the mean. Its error decreased rapidly at first, and then more slowly, to .30. The reduction in the error of the ability estimate is represented in Figure 19 by the decreasing width of the dotted line plotted around the ability estimate. The convergence of the ability estimate can be seen in the vertical plot of X's and 0's. Seventeen items were administered before the procedure reached the termination criterion of .30 for the standard error of the ability estimate.

Advantages and limitations. The primary advantage of the Bayesian testing methods is that they permit the tester to control the size of the error of measurement associated with the ability estimate obtained from any testee. This is

Figure 19
Report on a Bayesian Test

ID NUMBER: DATE TESTED: 73/05/18

X=CORRECT 0=INCORRECT ?=NO RESPONSE
ERROR BAND PLOTTED IS + AND - STANDARD DEVIATION

STAGE	LOW ABILITY LEVEL HIGH -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0	POSTERIOR ABILITY EST	POSTERIOR ABILITY SD
0	.....E.....	-.85	1.87
1	.....X.....	.19	1.30
2	.....0.....	-.88	.94
3	.....X.....	-.36	.76
4	.....0.....	-.73	.68
5	.....0.....	-1.07	.62
6	.....0.....	-1.89	.50
7	....X.....	-1.42	.44
8	...X...	-1.29	.40
9	...X...	-1.20	.39
10	...0...	-1.30	.36
11	...0...	-1.40	.36
12	...X...	-1.31	.35
13	...X...	-1.24	.33
14	...X...	-1.17	.33
15	...0...	-1.23	.32
16	...0...	-1.28	.32
17	...0...	-1.35	.30

17 ITEMS WERE ADMINISTERED

accomplished by continuing to administer items until the error of the ability estimate is reduced to a specified level. While this is an advantage of the model, it is possible that, given an actual application of the method with real items, termination based on this criterion might not occur for some testees.

In theory, the Bayesian testing methods appear to provide maximum adaptation to individual differences in ability level. In the absence of a pre-specified step size or offset, the Bayesian methods individualize the branching rule so that the next item to be administered is chosen to provide maximum information on each testee, based on the testee's unique sequence of responses.

The Bayesian testing methods function most effectively when the available item pool is very large and when items are highly discriminating. In Bayesian adaptive testing, most information is available from a given item response when the difficulty of the item exactly matches the estimated ability of the testee. To the extent that there is no item exactly at the testee's estimated ability level, the efficiency of the Bayesian procedures will be reduced somewhat, even though the next item selected for administration will be the "best" item from those that are available. A possible limitation of the Bayesian procedures, then, is that in actual testing, very large item pools with highly discriminating items distributed at closely spaced intervals throughout a wide ability range, will be required for maximum efficiency, and to insure that a majority of testees will reach the termination criterion.

A second limitation of the Bayesian procedures is that they require the use of fast computers, programmable in mathematically-related languages, to perform the complex calculations required after each item response to locate the next item and to revise the testee's ability estimate. This contrasts with all of the previously discussed adaptive testing methods, which could be administered by a relatively simple testing machine or an unsophisticated computer system programmed in simple, non-mathematically-based languages.

Research issues. A primary research question to be answered concerning the Bayesian strategies is the relative effectiveness and utility of the two somewhat different strategies proposed for Bayesian adaptive testing. Another major issue concerns the effect of violations of the assumptions on the results derived from the methods, since the Bayesian methods assume that certain distributions are normal. No data are as yet available concerning the effects of deviations of these distributions from normal on the efficiency of the Bayesian estimation procedures.

Research also needs to be conducted on termination rules in Bayesian testing. Since the number of items is usually free

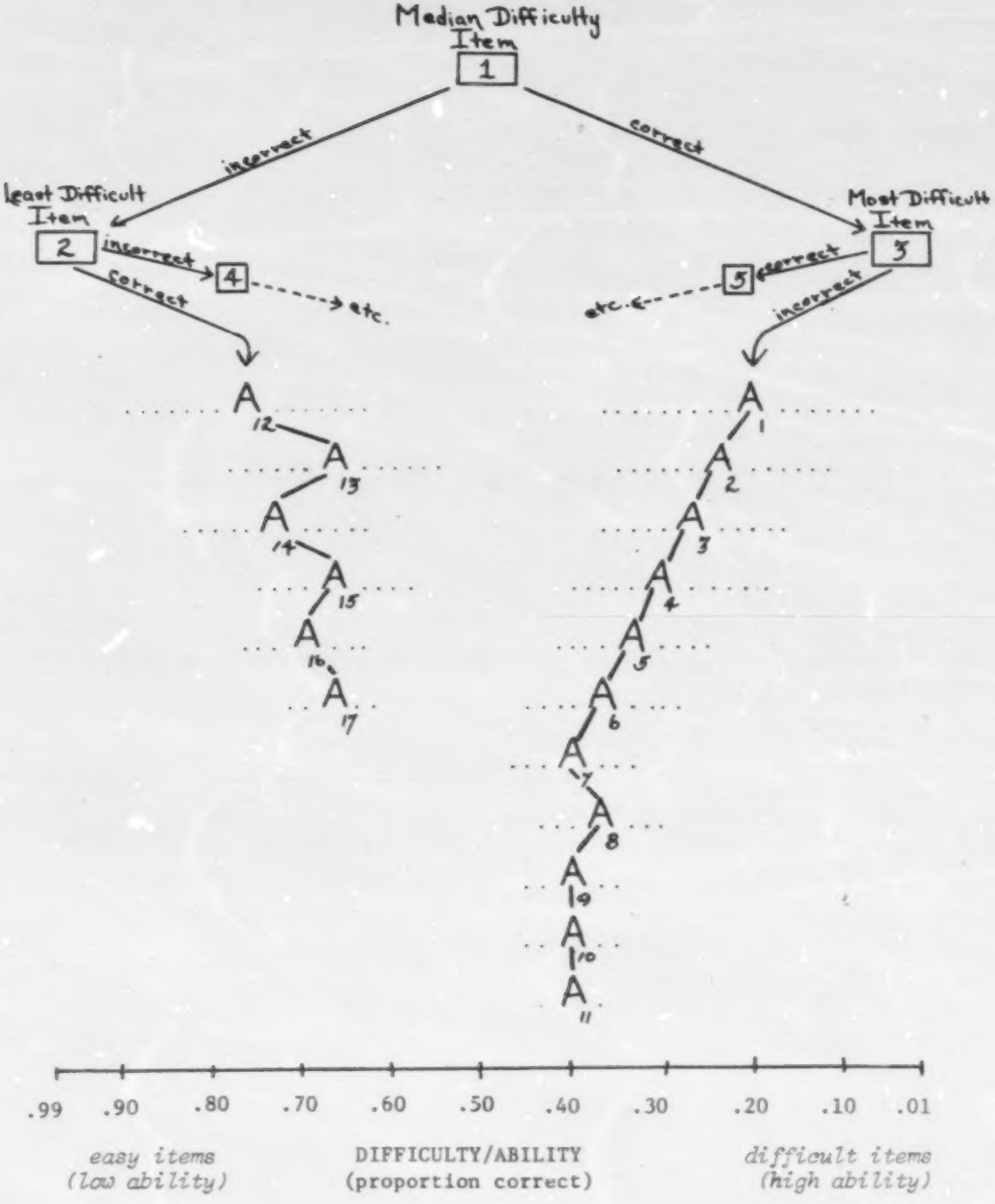
to vary, termination can occur on the basis of a pre-specified standard error of the ability estimate, as illustrated above. However, no data are available on realistic values of this standard error with real items and real testees. Similarly, no data are available on the effects of terminating testing after a fixed number of items. Another termination criterion which seems appropriate would be based on the analysis of changes in successive ability estimates from stage to stage. Such an index might be based on the reduction of changes in the ability estimates themselves (rather than their standard errors) to some pre-specified value. Also in need of investigation is the longitudinal stability of ability estimates derived from use of the different termination criteria.

Maximum Likelihood Strategies

Urry (1970) developed an adaptive testing strategy based on the maximum likelihood methods of modern test theory. The maximum likelihood procedure operates in a fashion similar to that of the Bayesian procedures, although the mathematical rationale is quite different. After a testee has answered one item correctly and another incorrectly, it is possible to solve maximum likelihood equations and obtain an ability estimate and its standard error. The next item selected for administration is the item in the item pool which has a difficulty level closest to the testee's estimated ability level. That item is administered and the testee's response is evaluated. Based on the testee's total item response record, now including the last item administered, the maximum likelihood equations are then solved again and a new ability estimate and its standard error are obtained. If the last item was answered correctly the testee's ability estimate will be somewhat higher; if the last item was incorrect, his ability estimate will be lower. At each stage of the procedure, the standard error of the ability estimate decreases as a result of the new information obtained from the last test item, in conjunction with his response pattern on all previous items.

The method begins using a partially structure item pool, until an initial ability estimate can be made. There is, then, a defined pathway that each testee follows until he answers at least one item correctly and one item incorrectly. After that criterion is met, the maximum likelihood estimation procedures begin, and a fixed branching procedure no longer occurs. The maximum likelihood procedure yields an ability estimate and its standard error, following a testee's response to each item. Testing can be terminated when the standard error of the ability estimate reaches a predetermined value. A diagrammatic representation of the method is shown in Figure 20. The first item administered to every testee is an item of average difficulty. If the testee answers the first item correctly, he is then administered the most difficult item in the item pool; if he answers the first item incorrectly, he is administered the

Figure 20
Hypothetical Response Records from Urry's Maximum Likelihood Adaptive Testing Strategy



easiest item in the pool. This procedure is necessary in order to solve the maximum likelihood equations. A correct answer to item 1 means simply that the testee's ability is likely to be above the median, but whether his ability is a z=0.01 or z=4.00 cannot be determined from just one item response. Similarly, an incorrect answer to item 1 indicates that the testee's ability is probably below the median--but it could be as high as z=-.05 or as low as z=-3.75, for example.

If the testee answers item 3 incorrectly, the maximum likelihood estimation procedure yields an initial estimate of the testee's ability, depicted in Figure 20 as A_1. The ability estimate, A_1, is between the most difficult item and the item of median difficulty, with a large error of estimate shown by the dots on either side of A_1.

In a similar fashion, the testee who answers item 1 incorrectly is administered the easiest item in the pool. If he answers that item (no. 2) correctly, it is possible to estimate that his ability lies between the difficulties of the easiest item and the item of median difficulty. The maximum likelihood estimate A_{12} and its standard errors are then computed.

For the high ability testee who has answered both items 1 and 3 correctly, the next most difficult item available is administered (item 5) in an attempt to obtain sufficient information to begin maximum likelihood estimation. Should that item also be answered correctly, it is followed by each next most difficult item in succession, until an incorrect response is encountered; at that point the maximum likelihood estimation procedure begins. Conversely, for the testee who answers items 2 and 4 incorrectly, additional items increasing in difficulty are administered until a correct response is obtained; at that point maximum likelihood estimation begins.

The examples shown in Figure 20 indicate that the item administered for the testee at A_1 will not be the same item administered to the testee whose ability is at A_{12}. In each case, the item chosen from the pool is the one closest in difficulty to the respective ability levels. A correct response to an item leads to a higher ability estimate (e.g., A_{13}) with a smaller error, while an incorrect response leads to a lower ability estimate (e.g., A_{14}) also with a smaller error. The procedure operates essentially as a decreasing step size procedure as it converges on the testee's ability level, with the step size determined on an intra-individual basis from the testee's total response pattern at each stage of testing.

Reckase (1973, 1974) has developed a maximum likelihood strategy which differs in several respects from Urry's. While Urry begins all testees with an item of average difficulty, Reckase allows different entry points into the item pool based on prior information about the testee's ability. If an estimate of the

testee's ability is available prior to the administration of the first test item, Reckase begins testing with the item in the pool which has difficulty closest to the testee's estimated ability.

Reckase also uses a different procedure for obtaining the data necessary to begin the maximum likelihood estimation procedure. As indicated above, maximum likelihood estimation of ability cannot begin until the testee's response record includes a correct and an incorrect response. In Reckase's strategy, a correct response to the first item administered leads to a stage 2 item twice as difficult as the first (using the Rasch (1966a,b) item easiness parameter). If that item is answered correctly, the next item administered is again twice as difficult. This process continues until an incorrect answer is obtained. Similarly, if the first item is answered incorrectly, a correct response is needed to begin maximum likelihood ability estimation. In this case, the second and succeeding items answered incorrectly are each half as difficult as the preceding item. Item administration continues in this fashion until a correct response is obtained.

Once maximum likelihood estimates of ability are available from the data in the testee's response record, Reckase's approach again differs slightly from Urry's. The objective of both approaches is to find and administer that item remaining in the pool which is closest to the testee's ability level. However, in a finite pool of real items there is not likely to be an item with difficulty exactly equal to the estimated ability of the testee. Urry resolves this problem by selecting the item closest in difficulty to the ideal item; thus, some selected items will be slightly easier than desired and other items will be slightly more difficult. Reckase, on the other hand, always chooses that less difficult item which is closest in difficulty to the ideal item, even though there may be a more difficult item which is closer in difficulty to the ideal item. This procedure might serve to reduce guessing effects somewhat in comparison to Urry's procedure.

Since Reckase's strategy can make use of prior information it might result in savings of a few items in comparison to Urry's. On the other hand, Urry's procedure will usually begin the maximum likelihood estimation process more quickly than Reckase's, resulting in a savings in testing time. While Urry's procedure for choosing non-ideal items might result in some guessing on items of higher difficulty than desired, Reckase's procedure might result in the administration of a set of items providing less average information per item than Urry's. The result would be more items administered in Reckase's than Urry's procedure.

Reckase's strategy does not explicitly take account of guessing, while Urry's model does. In Urry's model, test items

can vary in discrimination while Reckase assumes that his items all have equal discriminations. Thus, Reckases's model, as currently implemented, permits items to differ only in terms of their difficulties.

Both the maximum likelihood strategies and the Bayesian strategies effect "maximum" reduction in the error of the ability estimate at each stage of testing, although that reduction is achieved differently for the two methods. The Bayesian methods utilize normal distribution assumptions, although the maximum likelihood methods do not. Owen's (1969) Bayesian method derives each new ability estimate from the prior ability estimate in conjunction with the item response data on each new test item. On the other hand, the maximum likelihood methods re-estimate ability at each stage from the entire pattern of item responses, including the new item. Thus, the maximum likelihood methods give results which do not depend on the order in which items are administered, while ability estimates resulting from the Bayesian methods might have some dependence on the order in which items are administered to a given testee.

Advantages and limitations. Logically, the maximum likelihood strategies appear to be important competitors to the Bayesian strategies in terms of efficiency of testing time, since each item administered to a testee is chosen to provide maximum information on the testee's ability level. The methods also, like the Bayesian methods, provide individualized decreasing step size, an individualized number of items, and the capability of controlling the precision of the final ability estimate within the limitations of the item pool characteristics. These methods appear to have two primary limitations. First, it might be difficult to implement the maximum likelihood methods as operational testing strategies because of the time-consuming and complex calculations involved in deriving ability estimates. A second limitation lies in the fact that the maximum likelihood estimation procedure cannot begin until a mixed response pattern (some items correct and some incorrect) can be obtained. This last limitation means that ability estimates cannot be derived for individuals who answer all items correctly or all items incorrectly. This limitation is not characteristic of the Bayesian methods, although those methods will quickly run out of items for very high or low ability testees in most real item pools, and therefore fail to reach a termination criterion based on the error of the ability estimate.

Urry's maximum likelihood method also appears to have some problem in recovering from chance successes due to guessing at the early stages of testing. Chance successes due to guessing would appear to be most serious on the first item, since they would result in ability estimates quite divergent from the testee's real ability. As a result, it might take substantial numbers of items to locate the appropriate level of item difficulty for the testee whose early ability estimates are inflated

by chance successes. Reckase's method would seem to be less susceptible to this problem, because the second and succeeding items are not as extreme in difficulty as in Urry's strategy.

Research issues. Because Urry's (1970) simulation study and Reckase's (1974) empirical study based on 17 testees are the only ones which have used the maximum likelihood estimation strategies, very little is known about their characteristics. Like the Bayesian strategies, it would be important to know how sensitive results from these strategies are to deviations from the assumptions underlying the mathematical procedures. Similarly, research needs to be conducted to clarify termination rules, to study the effect on speed of termination and accuracy of final ability estimates of variations in composition of the item pool, and to determine their utility in actual computerized testing. This latter characteristic might be important in that the iterative calculations required for solution of the maximum likelihood equations at each stage of testing might result in substantially longer times between each item presentation than for other adaptive testing strategies.

EVALUATION

The major potential advantage of adaptive testing is that ability estimates derived from the use of adaptive strategies will have equal precision throughout the range of measured ability, and therefore provide scores of higher reliability and validity. In more traditional psychometric language, adaptive testing has the potential to equalize the error of measurement for all ability test scores. This is in contrast to the typical conventional test in which the most accurate measurement (smallest error of measurement) is for testees of mean ability, and the error of measurement generally gets larger as ability deviates from the mean. In other words, in the language of modern test theory (Lord & Novick, 1968), the information function of the conventional peaked test will approximate a normal distribution, while the adaptive test's information function approaches a horizontal line (Lord 1970, 1971a,c,d,e).

Adaptive testing achieves its equi-precision by locating a region of the item pool which is maximally appropriate for each testee. Items in that region of the item pool are those closest to the individual's estimated ability level. Each of these items will have a probability of being answered correctly of about .50. As a result, each item will provide maximum information on each testee's ability. At the same time, the more items there are administered, each of which provides near maximum information on the testee's ability, the more precise will be the resulting final ability estimate.

Each strategy of adaptive testing adapts item administration to the testee's ability level differently. Thus, it is possible to evaluate these methods using the criteria just described. That is, the strategies of adaptive testing can be compared in terms of how well the adaptive procedure appears to locate a region of the item pool which will provide maximum information per item for testees of various ability levels. In the absence of empirical data on each strategy, such an evaluation should help identify those adaptive strategies most appropriate for empirical research.

While each strategy will be ranked below on its potential to provide equi-precise scores, other characteristics of the strategies will be taken into account in the evaluation. One such characteristic is each strategy's use of non-test prior information on the ability level of the testee, which can be helpful in reducing testing time. Secondly, the strategies can be compared in terms of their susceptibility to guessing. A closely related characteristic is their capability of recovering from errors in routing, whether the errors be due to guessing or to testee response errors. A final criterion is the feasibility of implementing the strategies, since there are obviously wide differences among the strategies in this respect. Other evaluative criteria, unique to certain strategies, will be integrated into the rankings as they appear to be appropriate.

Bayesian and maximum likelihood strategies. In general, the variable branching strategies appear to provide the best potential for achieving equi-precise scores because they come the closest of all strategies to adapting the "step size" to individual differences in ability at each stage of testing. Essentially, these strategies attempt to find, at each stage of testing, the item in the item pool most nearly matched with the individual's estimated ability level at that stage in the sequential procedure. That item is administered, and the ability estimate is re-calculated before the next item is administered. These strategies also have the advantage that the number of items administered to each testee is not fixed in advance. Thus, testing can continue for most testees until the resulting ability estimate has a desired degree of precision. Since the degree of precision can be determined in advance, it is obvious that these methods can provide scores that are equally precise across the range of measured abilities.

It is difficult to determine on purely logical grounds whether the Bayesian or the maximum likelihood strategies will be more consistent in providing equi-precise scores with real item pools. The Bayesian methods and Reckase's maximum likelihood method have the advantage of utilizing different entry points into the item structure by taking into account prior information on each testee; this should help them reach convergence on the testee's ability level more rapidly. On the other hand, erroneous prior information can result in longer

testing times and biased ability estimates. It is also not clear whether the Bayesian methods will completely recover from wrong prior information, and if so how many items it will take in specific cases. On the positive side, Owen's Bayesian approach is designed to take account of guessing, using a model in which the selection of the next item to be administered varies as a function of the difference between the testee's estimated ability level and the difficulty of the last item answered.

Although the Bayesian strategies are designed to take into account information on the testee's ability level prior to the beginning of testing, Reckase's maximum likelihood strategy also permits variable entry points into the item structure. That method is limited, however, in that it is not currently designed to take account of guessing. The maximum likelihood methods yield ability estimates which are not dependent on order of item administration, since ability is re-estimated using the entire pattern of item responses after each item is administered. In the Bayesian approaches, the order of administration of the items might have an effect on ability estimates. For example, a chance success on an early item might result in Bayesian ability estimates which do not converge properly on a low ability estimate using a finite pool of real test items. Administration of the same item, and a resulting chance success, at a much later stage of testing will likely have less effect on the Bayesian ability estimate even given the same item pool. Should such item order effects be found in real item pools with the Bayesian approaches they will seriously reduce the potential utility of these strategies.

Although the maximum likelihood and Bayesian strategies rank highest in potential for equi-precise scores, they have some limitations that might affect their practical utility. First, the Bayesian methods require assumptions that certain distributions are normal. The appropriateness of this assumption and its robustness will require careful study. A second limitation of both methods is that their implementation requires very large and carefully calibrated item pools. Since these methods function by finding, at each stage of testing, the item most closely matched to the testee's ability level, they must have available a number of items at or near all potential ability levels. No data is yet available on how these methods function with real, finite, item pools on testee groups of wide-ranging ability. A final limitation of these methods is a uniquely practical one. Both the Bayesian and maximum likelihood methods are based on the solution of complex mathematical equations following each item response. Consequently, these strategies of adaptive testing are feasible only under computer administration. Not only are computers required for administering tests by these strategies, but they require fast computers programmable in mathematically-based languages in order to complete the computations with no noticeable delay in the interactive testing situation.

The stradaptive test. Among the fixed branching models, the stradaptive test appears to provide the best potential for equi-precise measurement. This results from the fact that it is structured so that there are an approximately equal number of items relevant to each of a number of different levels of ability. As a result, there should be sufficient items near the ability of most testees to provide a nearly constant level of precision for most ability levels. The branching procedure is designed to converge upon a testee's ability level, and to administer a series of items near that ability level to obtain a stable estimate of ability. The stradaptive test also has some of the advantages of the variable branching models. As do the Bayesian and maximum likelihood strategies, the stradaptive test permits an individualized number of items to be administered to each testee. Consequently, it too has the potential to control the degree of accuracy of a test score and to allow the calculation of an individualized "error of measurement." Further research is needed, however, on exactly how best to measure the precision of an ability estimate within the stradaptive test. This strategy also permits the use of variable entry points, with consequent reduction in testing time. In contrast to the Bayesian and maximum likelihood approaches, however, the stradaptive test quickly recovers from erroneous prior information. At the same time, the stradaptive test is explicitly designed to take account of chance successes resulting from guessing. A further advantage of the stradaptive test is that it does not require any assumptions about the distribution of ability in the population being measured.

The stradaptive test can be administered by a specially designed testing machine or by a computer. When computer-administered, however, only minimal arithmetic computations are involved, primarily in determining final scores. Consequently, relatively slow computers programmed in unsophisticated languages could be used for on-line administration of stradaptive tests.

Further research is needed to determine the optimal number of strata in the stradaptive item pool, and the minimum number of items necessary for each stratum. However, it appears that the stradaptive test makes more efficient use of its items than do the other fixed branching strategies and requires a smaller item pool than the variable branching strategies.

Truncated pyramids. Closely related to the stradaptive strategy are the truncated pyramids, using reflecting and retaining barriers. As shown in Figure 9, the item structures for these tests result in a number of items at each of a number of difficulty levels (the "strata" of the stradaptive test). If truncated pyramids had a large number of stages (e.g., 15) and had items at, say, eleven difficulty levels, their scores might be more nearly equi-precise. Also in their favor is their ability to recover from mis-routings relatively quickly,

if the pyramid has enough stages. The truncated pyramids have the disadvantage, however, that the number of items administered to each testee is constant. Since some individuals would be less consistent in their interactions with a given item pool, their scores will likely be more unreliable. As a result, fixing in advance the number of items to be administered to every testee will result in a tendency for the scores to be less equi-precise than if the number of items is permitted to vary until a given degree of precision is reached.

Two-stage tests. The two-stage test would appear to result in relatively equi-precise scores if there were a relatively large number of measurement tests well distributed across the ability continuum. While research to date has considered two-stage tests with only four or five measurement tests, a two-stage test using nine or ten measurement tests would be an important competitor to the models ranked above it. Under these circumstances the measurement tests could probably consist of as few as fifteen items and still obtain relatively accurate measurement at all ability levels. A two-stage test of this type would be similar in item structure to the truncated pyramid or the stradaptive test. The difference, however, is in the branching rule, which is the major drawback to the two-stage test.

Because the two-stage test uses only one branching decision, mis-routings will result in lower accuracy of measurement for some testees. For example, guessing on the routing test might lead to the assignment of some testees to measurement tests that are too difficult (i.e., not appropriate for their ability level). Similarly, chance or inattention on the routing test could lead to the assignment of other testees to measurement tests that are too easy. Such misrouting might be corrected for in two ways. First, computer administration could be used in which the testee's responses to the first few items in the measurement test are carefully monitored. Should the testee answer, say, the first five items all correctly or incorrectly, he could then be re-routed to a more difficult or easier measurement test, respectively. Under this mode of administration, the two-stage test would take on some of the chacteristics of the stradaptive test and the multiple-item pyramids, using obviously different branching rules.

A second way of partially correcting for mis-routings is to use a double-routing test, as suggested by Cleary, et al., (1968) and shown in Figure 3. Double routing gives the testee a chance to partially recover from routing errors on the first routing test, but not on the second. The double-routing procedures, however, share the other deficiences of two-stage tests, which serve to lower their precision of measurement at some ability levels. Primary among these is the fact that the number of items administered to all testees is constant. Furthermore, two-stage tests do not make use of prior information on the testee's ability. On the positive side, two-stage tests

are practical since they can be administered by paper and pencil, and they do not require very large numbers of items for their construction.

Multiple-item pyramids. Next in their apparent potential for providing equi-precise measurement are the pyramidal structures with several items per stage, as illustrated in Figure 10. These multiple-item pyramids are the natural extension of the double-routing two-stage test to a test combining multiple routing with simultaneous measurement tests. Because these tests have fairly large numbers of items at each of a number of difficulty levels, they will likely provide scores of higher precision than the lower ranked models. They share a major disadvantage of all pyramidal models in that for testees at the higher and lower ability levels most items are used in routing with few items available for actually measuring the testee's ability level. Also in common with all pyramidal models, the multiple-item pyramid is not completely recoverable from mis-routings at the early stages for testees of high or low ability.

Other pyramidal models. The remainder of the pyramidal models appears to provide less potential for equi-precise measurement. Within this general group, however, the remaining models appear to fit into a hierarchy. The Robbins-Monro pyramids appear to provide promise of relatively equi-precise measurement. However, empirical research on them is generally inappropriate since they are practically unfeasible due to the large numbers of items they require. Robbins-Monro pyramids also appear to be inappropriate for use with multiple-choice tests where guessing is possible since they do not appear to provide complete recovery from chance successes. Paterson's (1962) decreasing step size pyramid ranks highest of the practically feasible methods since its items are more nearly rectangularly distributed across the potential ability range than are the items in competing models. Thus, Paterson's approach will result in a pyramidal structure with a small but relatively constant number of items at each of a number of regions throughout the range of abilities. Although the differential option branching pyramids require a relatively complex item branching network, they use more of the information in a testee's response and should thus provide relatively precise measurement. Finally, the standard pyramids (e.g., Figure 4) appear to be the least likely of the pyramidal strategies to provide good measurement charcteristics. On a comparative basis, they have fewer items at or near each ability level, with too few items at the extremes to measure with equal precision across the ability range. Furthermore, the basic pyramidal structure does not allow complete recovery from branching errors for some testees, nor does it permit the number of items administered to a testee to be varied to control the magnitude of the errors of measurement.

The flexilevel test. For a number of reasons, the flexilevel test appears to provide the least promise of the methods reviewed for equi-precise measurement. First, the flexilevel structure provides only one item at each of a number of ability levels. As a result, only a few items will provide maximum information for each testee. Secondly, the divergence procedure followed in the flexilevel test operates contrary to the goal of locating an area of the item pool which will provide maximum information on the testee's ability level. For testees of average ability, the item difficulties continually diverge from the testee's ability level at successive stages of testing; each item therefore, provides less and less information. For testees of high and low ability, the early items in the test are used to route to items at the testee's ability level, then item difficulties diverge again from that ability level; the result, again, is the administration of a series of items providing less and less information. An additional effect of this divergence procedure is that it might also result in random guessing near the end of the test as alternate items become much too difficult. This major disadvantage of the flexilevel test, however, results primarily from the use of a fixed termination criterion. Should the flexilevel procedure be re-designed to detect divergence as it begins to operate and to terminate the test at that point, thereby allowing the number of items administered to a testee to vary, it might provide more equi-precise measurement. However, because it has only one item at each of a number of ability levels, the number of items administered to each testee will be quite small, with only a few items providing any information on the testee's ability. The result will be scores that are less precise, and therefore, less reliable, than those derivable from other methods of adaptive testing.

Advantages of the flexilevel procedure include its potential for paper and pencil administration, and its small item pool. To administer a flexilevel test by paper and pencil, however, requires an answer sheet which informs the testee of the correctness of his response to each item, and a testee motivated to and capable of following the routing instructions.

Summary. Based on their capability of providing ability estimates of equal precision throughout the ability range, the Bayesian and maximum likelihood strategies of adaptive testing appear to hold the most promise for future application. However, these methods require the availability of very large item pools and computers for their administration. Because the stradaptive test requires only automated administration and smaller item pools, yet has the capability to approximate the equi-precise measurement of the first two strategies, it appears to be an important competitor as a feasible testing strategy for many applied situations. Ranked next in their capability to provide equi-precise measurement were the truncated pyramids, a recoverable two-stage test, double-branching two-stage tests, and the

multiple-item pyramidal models. The remainder of the pyramidal models appeared to rank below the multiple-item pyramids, with decreasing step size pyramids the most likely candidates for providing equi-precise measurement, followed by differential option branching pyramids and standard pyramids. Finally the flexilevel test appears to provide the least potential for measurement of equal precision throughout the range of ability due to its tendency to diverge from items of appropriate difficulty once they are encountered.

Very little is known either theoretically or empirically about any of the adaptive testing strategies described above. Obviously, a complete evaluation of the psychometric characteristics and practical utility of each of these strategies of adaptive testing will rest on an accumulation of research data.

References

Angoff, W.H. & Huddleston, E.M. The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Test. Princeton, New Jersey, Educational Testing Service, Statistical Report SR-58-21, 1958.

Bayroff, A.G. & Seeley, L.C. An exploratory study of branching tests. U.S. Army Behavioral Science Research Laboratory, Technical Research Note 188, June 1967.

Bayroff, A.G., Thomas, J.J. & Anderson, A.A. Construction of an experimental sequential item test. Research memorandum 60-1, Personnel Research Branch, Department of the Army, January 1960.

Betz, N.E. & Weiss, D.J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973. (AD 768993)

Betz, N.E. & Weiss, D.J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD A001230)

Cleary, T.A., Linn, R.L. & Rock, D.A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360.

Cory, C.H. An evaluation of computerized tests as predictors of job performance in three Navy ratings: I. Development of the instruments. Technical Report 75-2, Naval Personnel Research and Development Center, San Diego, California, August 1974.

Cowden, D.J. An application of sequential sampling to testing students. Journal of the American Statistical Association, 1946, 41, 547-556.

Cronbach, L.J. & Gleser, G.C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

DeWitt, L.J. & Weiss, D.J. A computer software system for adaptive ability measurement. Research Report 74-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD 773961)

Green B.F. Jr. Comments on tailored testing. In W.H. Holtzman (Ed.), Computer-assisted instruction, testing and guidance. New York: Harper and Row, 1970.

Hansen, D.N. An investigation of computer-based science testing. In R.C. Atkinson and H.A. Wilson (Eds.), Computer-assisted instruction: a book of readings. New York: Academic Press, 1969.

Hick, W.E. Information theory and intelligence tests. British Journal of Psychology, Statistical Section, 1951, 4, 157-164.

Krathwohl, D.R. & Huyser, R.J. The sequential item test (SIT). American Psychologist, 1956, 2, 419.

Larkin, K.C. & Weiss, D.J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD 783553)

Lindley, D.V. Introduction to probability and statistics from a Bayesian viewpoint. Cambridge, England: Cambridge University Press, 1965.

Linn, R.L., Rock, D.A. & Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Lord, F.M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance, New York: Harper and Row, 1970.

Lord, F.M. Robbins-Munro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31. (a)

Lord, F.M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151. (b)

Lord, F.M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711. (c)

Lord, F.M. A Theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813. (d)

Lord, F.M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-241. (e)

Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McBride, J.R. & Weiss, D.J. A word knowledge item pool for adaptive ability measurement. Research Report 74-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD 781894)

Moonan, W.J. Some empirical aspects of the sequential analysis technique as applied to an achievement examination. Journal of Experimental Education, 1950, 18, 195-207.

Mussio, J.J. A modification to Lord's model for tailored tests. Unpublished doctoral dissertation, University of Toronto, 1973.

Novick, M.R. Bayesian methods in psychological testing. Princeton, N.J.: Educational Testing Serice, Research Bulletin RB-69-31, 1969.

Owen, R.J. A Bayesian approach to tailored testing. Princeton, N.J.: Educational Testing Service, Research Bulletin RB-69-92, 1969.

Paterson, J.J. An evaluation of the sequential method of psychological testing. Unpublished doctoral dissertataion, Michigan State University, 1962.

Rasch, G. An individualistic approach to item analysis. In P.F.Lazarsfeld & N.W. Henry (Eds.), Readings in mathematical social science. Chicago: Science Research Associates, 1966. (a)

Rasch, G. An item analysis that takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57. (b)

Reckase, M.D. An interactive computer program for tailored testing based on the one-parameter logistic model. Paper presented at the National Conference on the Use of On-line Computers in Psychology, St. Louis, Missouri, 1973.

Reckase, M.D. An application of the Rasch simple logistic model to tailored testing. Paper presented at the Annual Meeting of the American Educational Research Association, 1974.

Schmitt, S.A. Measuring uncertainty: an elementary introduction to Bayesian statistics. Reading, Mass.: Addison-Wesley, 1969.

Stocking, M. Short tailored tests. Princeton, N.J.: Educational Testing Service, Research Bulletin RB-69-63, 1969.

Urry, V.W. A monte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Wald, A. Sequential Analysis, New York: Wiley, 1947.

Weiss, D.J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973. (AD 768376)

Weiss, D.J. & Betz, N.E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973. (AD 757788)

Winkler, R.L. Introduction to Bayesian inference and decision. New York: Holt, Rhinehart & Winston, 1972.

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.

DISTRIBUTION LIST

Navy

4 Dr. Marshall J. Farr, Director
Personnel and Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: Research Psychologist

1 ONR Branch Office
1030 East Green Street
Pasadena, CA 91101
ATTN: E.E. Gloye

1 ONR Branch Office
536 South Clark Street
Chicago, IL 60605
ATTN: M.A. Bertin

1 Office of Naval Research
Area Office
207 West 24th Street
New York, NY 10011

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

12 Defense Documentation Center
Cameron Station, Building 5
5010 Duke Street
Alexandria, VA 22314

1 Special Assistant for Manpower
OASN (M&RA)
Pentagon, Room 4E794
Washington, DC 20350

1 LCDR Charles J. Theisen, Jr., MSC, USN
4024
Naval Air Development Center
Warminster, PA 18974

1 Chief of Naval Reserve
Code 3055
New Orleans, LA 70146

1 Dr. Lee Miller
Naval Air Systems Command
AIR-413E
Washington, DC 20361

1 CAPT John F. Riley, USN
Commanding Officer
U.S. Naval Amphibious School
Coronado, CA 92155

1 Chief
Bureau of Medicine & Surgery
Research Division (Code 713)
Washington, DC 20372

1 Chairman
Behavioral Science Department
Naval Command & Management Division
U.S. Naval Academy
Luce Hall
Annapolis, MD 21402

1 Chief of Naval Education & Training
Naval Air Station
Pensacola, FL 32508
ATTN: CAPT Bruce Stone, USN

1 Mr. Arnold Rubinstein
Naval Material Command (NAVMAT 03424)
Room 820, Crystal Plaza #6
Washington, DC 20360

1 Director, Navy Occupational Task
Analysis Program (NOTAP)
Navy Personnel Program Support
Activity
Building 1304, Bolling AFB
Washington, DC 20336

1 Dr. Richard J. Niehaus
Office of Civilian Manpower Management
Code 06A
Washington, DC 20390

1 Department of the Navy
Office of Civilian Manpower Management
Code 263
Washington, DC 20390

1 Chief of Naval Operations (OP-987E)
Department of the Navy
Washington, DC 20350

1 Superintendent
Naval Postgraduate School
Monterey, CA 93940
ATTN: Library (Code 2124)

1 Commander, Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203
ATTN: Code 015

1 Mr. George N. Graine
Naval Ship Systems Command
SHIPS 047C12
Washington, DC 20362

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. Norman J. Kerr

1 Dr. William L. Maloy
Principal Civilian Advisor
for Education & Training
Naval Training Command, Code 01A
Pensacola, FL 32508

1 Dr. Alfred F. Smode, Staff Consultant
Training Analysis & Evaluation Group
Naval Training Equipment Center
Code N-00T
Orlando, FL 32813

1 Dr. Hanns H. Wolff
Technical Director (Code N-2)
Naval Training Equipment Center
Orlando, FL 32813

1 Chief of Naval Training Support
Code N-21
Building 45
Naval Air Station
Pensacola, FL 32508

1 Dr. Bernard Rimland
Navy Personnel R&D Center
San Diego, CA 92152

5 Navy Personnel R&D Center
San Diego, CA 92152
ATTN: Code 10

1 D. M. Gragg, CAPT, MC, USN
Head, Educational Programs Development
Department
Naval Health Sciences Education and
Training Command
Bethesda, MD 20014

Army

1 Headquarters
U.S. Army Administration Center
Personnel Administration Combat
Development Activity
ATCP-HRO
Ft. Benjamin Harrison, IN 46249

1 Armed Forces Staff College
Norfolk, VA 23511
ATTN: Library

1 Commandant
United States Army Infantry School
ATTN: ATSH-DET
Fort Benning, GA 31905

1 Deputy Commander
U.S. Army Institute of Administration
Fort Benjamin Harrison, IN 46216
ATTN: EA

1 Dr. Stanley L. Cohen
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph Dusek
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 Mr. Edmund F. Fuchs
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA

1 Dr. J.E. Uhlaner, Technical Director
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA 22209

1 HQ USAREUR & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

1 Major P.J. DeLeo
Instructional Technology Branch
AF Human Resources Laboratory
Lowry AFB, CO 80230

1 AFOSR/NL
1400 Wilson Boulevard
Arlington, VA 22209

1 Commandant
USAF School of Aerospace Medicine
Aeromedical Library (SUL-4)
Brooks AFB, TX 78235

Air Force

1 Research Branch
AF/DPMYAR
Randolph AFB, TX 78148

1 Dr. G.A. Eckstrand (AFHRL/AS)
Wright-Patterson AFB
Ohio 45433

1 AFHRL/DOJN
Stop #63
Lackland AFB, TX 78236

1 Dr. Robert A. Bottenberg (AFHRL/SM)
Stop #63
Lackland AFB, TX 78236

1 Dr. Martin Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

Marine Corps

1 Mr. E.A. Dover
Manpower Measurement Unit (Code MPI)
Arlington Annex, Room 2413
Arlington, VA 20380

1 Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Director, Office of Manpower Utilizat
Headquarters, Marine Corps (Code MPU)
MCB (Building 2009)
Quantico, VA 22134

1 Dr. A.L. Slafkosky
Scientific Advisor (Code RD-1)
Headquarters, U.S. Marine Corps
Washington, DC 20380

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-1/
U.S. Coast Guard Headquarters
Washington, DC 20590

Other DOD

1 Lt. Col. Henry L. Taylor, USAF
Military Assistant for Human Resources
OAD (E&LS) ODDR&E
Pentagon, Room 3D129
Washington, DC 20301

1 Col. Austin W. Kibler
Advanced Research Projects Agency
Human Resources Research Office
1400 Wilson Boulevard
Arlington, VA 22209

1 Helga L. Yeich
Advanced Research Projects Agency
Manpower Management Office
1400 Wilson Boulevard
Arlington, VA 22209

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. William Gorham, Director
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E. Street, N.W.
Washington, DC 20415

1 Dr. Andrew R. Molnar
Technological Innovations in Educati
National Science Foundation
Washington, DC 20550

1 Dr. Marshall S. Smith
Asst Acting Director
Program on Essential Skills
National Institute of Education
Brown Bldg, Rm 815
19th and M St., N.W.
Washington, D.C. 20208

Miscellaneous

1 Dr. Scarvia B. Anderson
Educational Testing Service
17 Executive Park Drive, N.E.
Atlanta, GA 30329

1 Dr. John Annett
The Open University
Milton Keynes
Buckinghamshire
ENGLAND

1 Dr. Richard C. Atkinson
Stanford University
Department of Psychology
Stanford, CA 94305

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

1 Dr. Bernard M. Bass
University of Rochester
Management Research Center
Rochester, NY 14627

1 Mr. Kenneth M. Bromberg
Manager - Washington Operations
Information Concepts, Inc.
1701 North Fort Myer Drive
Arlington, VA 22209

1 Centry Research Corporation
4113 Lee Highway
Arlington, VA 22207

1 Dr. Kenneth E. Clark
University of Rochester
College of Arts & Sciences
River Campus Station
Rochester, NY 14627

1 Dr. Rene' V. Dawis
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Norman R. Dixon
Room 170
190 Lothrop Street
Pittsburgh, PA 15260

1 Dr. Robert Dubin
University of California
Graduate School of Administration
Irvine, CA 92664

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 ERIC
Processing and Reference Facility
4833 Rugby Avnue
Bethesda, MD 20014

1 Dr. Victor Fields
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. Edwin A. Fleishman
American Institutes for Research
Foxhall Square
3301 New Mexico Avenue, N.W.
Washington, DC 20016

1 Dr. Robert Glaser, Director
University of Pittsburgh
Learning Research & Development Center
Pittsburgh, PA 15213

1 Mr. Harry H. Harman
Educational Testing Service
Princeton, NJ 08540

1 Dr. Richard S. Hatch
Decision Systems Associates, Inc.
11428 Rockville Pike
Rockville, MD 20852

1 Dr. M.D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO
Division No. 3
P.O. Box 5787
Presidio of Monterey, CA 93940

1 HumRRO
Division No. 4, Infantry
P.O. Box 2036
Fort Benning, GA 31905

1 HumRRO
Division No. 5, Air Defense
P.O. Box 6057
Fort Bliss, TX

1 HumRRO
Division No. 6, Library
P.O. Box 428
Fort Rucker, IL 36360

1 Dr. Lawrence B. Johnson
Lawrence Johnson & Associates, Inc.
200 S. Street, N.W., Suite 502
Washington, DC 20009

1 Dr. Milton S. Katz
MITRE Corporation
Westgate Research Center
McLean, VA 22101

1 Dr. Steven W. Keele
University of Oregon
Department of Psychology
Eugene, OR 97403

1 Dr. David Klahr
Carnegie-Mellon University
Department of Psychology
Pittsburgh, PA 15213

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Ernest J. McCormick
Purdue University
Department of Psychological Sciences
Lafayette, IN 47907

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Cortona Drive
Santa Barbara Research Park
Goleta, CA 93017

1 Mr. Edmond Marks
405 Old Main
Pennsylvania State University
University Park, PA 16802

1 Dr. Leo Munday, Vice-President
American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Donald A. Norman
University of California, San Diego
Center for Human Information Processing
LaJolla, CA 92037

1 Mr. Brian McNally
Educational Testing Service
Princeton, NJ 08540

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, VA 22207

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, CA 90265

1 Dr. Joseph W. Rigney
University of Southern California
Behavioral Technology Laboratories
3717 South Grand
Los Angeles, CA 90007

1 Dr. Leonard L. Rosenbaum, Chairman
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. George E. Rowland
Rowland and Company, Inc.
P.O. Box 61
Haddonfield, NJ 08033

1 Dr. Arthur I. Siegel
Applied Psychological Services
404 East Lancaster Avenue
Wayne, PA 19087

1 Dr. C. Harold Stone
1428 Virginia Avenue
Glendale, CA 91202

1 Mr. Dennis J. Sullivan
725 Benson Way
Thousand Oaks, CA 91360

1 Dr. Benton J. Underwood
Northwestern University
Department of Psychology
Evanston, IL 60201

1 Dr. David J. Weiss
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Anita West
Denver Research Institute
University of Denver
Denver, CO 80210

1 Dr. Kenneth N. Wexler
University of California
School of Social Sciences
Irvine, CA 92664

1 Carl R. Vest, Ph.D.
Battelle
Memorial Institute
Washington Operations
2030 - M Street, N.W.
Washington, D.C. 20036